UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Análise estatística de dados de metabolómica: identificação dos compostos envolvidos na resposta das plantas à simbiose com fungos ectomicorrízicos

Bruno de Campos Bento

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Prof.ª Doutora Lisete Maria Ribeiro de Sousa
Doutora Mónica Guita Sebastiana

2020

Agradecimentos

Primeiramente gostaria de agradecer às minhas duas orientadoras:

À Professora Lisete Sousa agradeço por todo o conhecimento estatístico que me proporcionou, por saber sempre como ajudar um aluno nos momentos em que este mais precisa e pela disponibilidade prestada.

À Doutora Mónica Sebastiana obrigado pela disponibilidade dos dados, pela sua simpatia, atenção e por todo o conhecimento biológico que me forneceu.

À minha família, obrigado por todo o apoio que me deram e por estarem sempre lá para mim, a vossa ajuda foi essencial.

Aos meus amigos e colegas de faculdade, um grande obrigado por todo o apoio. Sem vocês nunca teria conseguido atingir esta fase que outrora parecia completamente fora de alcance.

Um beijinho especial para a Raquel Fonseca, Rita Freire, Letícia Almeida, Heloísa Galante e Sofia Mateus por todo o apoio que me deram.

Por último a uma pessoa muito especial, ao Pedro Alves. O meu maior apoio e a minha pessoa favorita, o mundo é um lugar melhor porque existes nele. Obrigado por acreditares sempre em mim.

Um obrigado a todos,

Bruno Bento

03 de Março de 2020

Resumo

A associação simbiótica entre as raízes e fungos micorrízicos do solo é extremamente importante para a nutrição e desenvolvimento das plantas. Apesar de haver já algum conhecimento sobre os genes (genoma/transcritoma) e proteínas (proteoma) envolvidos no processo de micorrização, sabe-se muito pouco sobre as alterações dos metabolitos da planta hospedeira quando esta se encontra em simbioses com o fungo micorrízico. Assim, a identificação do conjunto de metabolitos ou compostos biológicos (metaboloma) envolvidos na micorrização reveste-se de especial interesse para completar a informação fornecida pelo genoma/transcritoma e proteoma. Estes são os últimos produtos da expressão dos genes e, por isso, são extremamente informativos acerca das alterações moleculares que estão a acontecer nos organismos vivos.

Sabe-se que os organismos vivos sujeitos a stresses ambientais (e.g. seca, salinidade do solo) ou durante interações com microrganismos simbióticos ou patogénicos, podem produzir metabolitos específicos ou sofrer alterações na quantidade de metabolitos já produzidos. A identificação destas alterações permite saber, a nível molecular, o que está a acontecer quando uma planta estabelece, por exemplo, uma relação simbiótica com um fungo da rizosfera.

Neste estudo pretendeu-se contribuir para a identificação das alterações no metaboloma do sobreiro quando este se encontra em simbiose com um fungo micorrízico do solo. O metaboloma da raíz micorrizada do sobreiro foi extraído usando quatro frações compostas por diferentes solventes: água, metanol, acetonitrilo e clorofórmio (orgânica). Foram extraídas seis réplicas de plantas de sobreiros – sendo três micorrizadas e três utilizadas como grupo de controlo não micorrizado, para efeitos de comparação.

Através da aplicação de métodos estatísticos tais como *Rank Products*, Modelos Lineares e Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA) , foi possível a identificação dos metabolitos diferencialmente acumulados que possibilitarão a descoberta de novas informações sobre a identidade dos metabolitos envolvidos na simbiose entre o sobreiro e o fungo micorrízico, promovendo novos conhecimentos sobre as vias metabólicas implicadas na micorrização.

Palavras-Chave: Ectomicorrizas; *Quercus suber L.*; Metabolitos; *Rank Products*; Análise discriminante por mínimos quadrados parciais.

Abstract

The symbiotic association between roots and mycorrhizal fungi is extremely important for plant nutrition and development. Although there is already some knowledge about the genes and proteins involved in the mycorrhization process, very little is known about changes in the host plant metabolome when in symbiosis with the mycorrhizal fungus. Thus, the identification of the set of metabolites or biological compounds involved in mycorrhization (metabolome) is of particular interest to supplement the information provided by the transcriptome (transcript) and proteome (protein). These are the latest products of gene expression and are therefore extremely informative about the molecular changes that are happening in living organisms.

It is well known that living organisms produce specific metabolites or suffer alterations in metabolite levels when subjected to environmental stresses (e.g. drought, soil salinity) or interactions with microorganisms, pathogenic or symbiotic. By identifying these changes, it is possible to know at a molecular level what is happening when a plant establishes a symbiotic relationship.

The aim of this study was to contribute for the identification of changes in the metabolome of the cork oak during symbiosis with a soil mycorrhizal fungus. The mycorrhizal root metabolome was extracted using four fractions composed of different solvents: water, methanol, acetonitrile and chloroform (organic). Six replicates of cork oak plants were extracted − three mycorrhizal and three used as non-mycorrhizal control group for comparison.

Through the application of statistical methods for linear models of microarray data, like Rank Products and Linear Models, and PLS-DA (*Partial Least Squares Discriminant Analysis)*, it was possible to identify the differentially accumulated metabolites that will allow the discovery of new information about the identity of the metabolites involved in the symbiosis between cork oak and the mycorrhizal fungus, allowing new knowledge about the metabolic pathways involved in mycorrhizal symbiosis.

Keywords: Ectomycorrhizal; *Quercus suber L.*; Metabolites; *Rank Products*; *Partial Least Squares Discriminant Analysis*.

Índice

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas, Siglas e Símbolos

ACN	Acetonitrilo
ESI	Ionização por electrospray
FC	*Fold-Change*
FTICR	*Fourier-Transform Ion-Cyclotron-Resonance*
MeOH	Metanol
PLS-DA	*Partial Least Squares Discriminant Analysis*
Read	Sequência curta de nucleotídeos
VIP	*Variable Importance in Projection*
VPN	Valor Preditivo Negativo

Capítulo 1

Introdução

Este trabalho insere-se num estudo da área da Biologia que pretende identificar e caracterizar o metaboloma da raíz do sobreiro quando esta se encontra em simbiose com um fungo benéfico do solo, estabelecendo uma associação denominada de micorrizas. Este estudo tem como objetivos principais a identificação dos compostos químicos presentes na raíz do sobreiro micorrizado, bem como a identificação das alterações nesses compostos induzidos pela micorrização, ou seja, identificar quais os compostos da raíz de sobreiro, induzidos ou reprimidos pela micorrização. Para isso foram comparadas plantas micorrizadas e não micorrizadas no sentido de se identificar os compostos diferencialmente acumulados, ativados ou reprimidos, sendo que os ativados se encontram em maior quantidade nas plantas micorrizadas do que nas não micorrizadas e os reprimidos ao contrário. Para identificação destes compostos foram aplicados vários métodos estatísticos que têm sido usados para analisar dados de experiências utilizando tecnologias das chamadas Ómicas (e.g. *microarrays* de DNA, sequenciação em massa de DNA).

As micorrizas são associações simbióticas estabelecidas entre as raízes das plantas terrestres e fungos especializados do solo. Esta relação simbiótica, um fenómeno generalizado que se estima ocorrer em cerca de 80% das espécies vegetais, evoluiu como uma adaptação das plantas ao ecossistema terrestre (Wang & Qiu, 2006), sendo essencial para a boa nutrição das mesmas e para a qualidade do solo (Smith & Read, 1997). Esta relação simbiótica, ao possibilitar uma troca mútua de nutrientes, revela-se reciprocamente benéfica – evidenciada no presente estudo, que registou observações de plantas de sobreiro e fungos ectomicorrízicos. Para uma análise precisa, foi utilizado um grupo de controlo não inoculado constituído por plantas que serviram de métrica comparativa com as suas homólogas inoculadas com o fungo micorrízico.

A metabolómica trata do estudo do metaboloma, isto é, do conjunto de todos os metabolitos de uma célula, tecido ou organismo; da mesma forma que a genómica examina o conjunto total de genes, e a proteómica examina o conjunto total de proteínas (Han et al. 2008). Os estudos de metabolómica têm como objetivo a identificação e quantificação dos metabolitos produzidos sob uma determinada condição experimental (Fiehn et al. 2001; Goodacre et al. 2004). Os metabolitos são os produtos finais da expressão genética e da função das proteínas, e incluem uma variedade diversificada de biomoléculas, como aminoácidos, lípidos, açúcares, antioxidantes, cofatores, hormonas, inibidores de enzimas, neurotransmissores etc. (Gamache et al. 2004). Quantificar as perturbações destes metabolitos em resposta a doenças específicas, drogas, modificações ambientais ou genéticas (Denkert et al. 2006; Goodacre et al. 2004; Nicholson et al. 1999) é extremamente importante para a compreensão das vias bioquímicas envolvidas nesses processos.

1.1 O Sobreiro e os fungos micorrízicos

O sobreiro (*Quercus suber L.*) é uma espécie perene, típica da região ocidental do Mediterrâneo. Em Portugal, o sobreiro é um importante recurso económico devido ao valor comercial da cortiça, sendo esta um produto natural originário da casca renovável do sobreiro e tendo aplicações polivalentes.

Na natureza, as raízes das plantas estão envolvidas numa associação simbiótica com fungos do solo, denominada por micorrizas. Estas associações naturais existem abundantemente nas florestas temperadas e boreais, onde fungos do solo colonizam as raízes de espécies de árvores dominantes, como os carvalhos, os pinheiros, as faias ou os choupos (Smith & Read, 1997).

Nesta associação, que remonta a 120 milhões de anos (Brundrett, 2002), os fungos micorrízicos transferem ativamente nutrientes e água para a planta hospedeira. Em troca, a planta pode transferir até 1/3 dos açúcares produzidos fotossinteticamente para os fungos (Nehls et al. 2007). Esta troca de nutrientes é essencial para a persistência de plantas e fungos, principalmente em solos pobres em nutrientes, sendo as micorrizas uma maneira de superar as limitações de nutrientes e carboidratos enfrentadas por árvores e fungos nos ecossistemas florestais (Nehls et al. 2007). Nas micorrizas das espécies florestais como o sobreiro, o micélio fúngico forma uma camada de hifas que se enrolam em redor das raízes mais curtas, isolando-as do solo circundante (Sebastiana et al. 2014). Estas micorrizas são chamadas de ectomicorrizas (Fig. 1.1).

Figura 1.1 - Raízes micorrizadas na planta do sobreiro (Créditos: Mónica Sebastiana)

Nas raízes micorrízicas, o fungo coloniza a raíz e ao mesmo tempo desenvolve um micélio extraradicular que se estende no solo e forma uma rede de hifas especializadas na aquisição de nutrientes e água, contribuindo dessa maneira para um aumento substancial da área de absorção radicular (Sebastiana et al. 2013). Além disso, a simbiose micorrízica beneficia a saúde das plantas, aumentando a sua proteção contra stresses bióticos (por exemplo, ataque por microrganismos patogénicos) e abióticos (por exemplo, seca, salinidade, contaminação do solo com metais pesados) e melhora a estrutura do solo, promovendo a sua agregação (Brazanti et al. 1999; Adriaensen et al. 2003; Barea et al. 2011). A simbiose ectomicorrízica é essencial para a vitalidade e saúde das

árvores em florestas temperadas e boreais, onde desempenha um papel importante no ciclo de nutrientes e no funcionamento dos ecossistemas florestais (Sebastiana et al. 2014).

O sobreiro é abundantemente produzido em viveiros para fins de regeneração florestal (Sebastiana et al., 2013). No entanto, devido aos elevados níveis de fertilizantes e pesticidas aplicado nos viveiros, a fim de apressar o seu crescimento inicial, os sistemas radiculares das plantas jovens são desprovidos de fungos micorrízicos simbióticos que favoreceriam o seu crescimento e sobrevivência pós-plantação, verificando-se altas taxas de mortalidade em novas plantações (Sebastiana et al. 2013). A produção de plantas jovens com ectomicorrizas desenvolvidas em viveiro é considerada uma estratégia promissora para minimizar o choque inicial do transplante e aumentar a sobrevivência e o crescimento das plantas durante os primeiros anos de uma plantação: a sua fase mais crítica (Sebastiana et al. 2013). A inoculação micorrízica pode melhorar não apenas o crescimento das plantas, mas também o seu *status* fisiológico, melhorando a sua capacidade fotossintética e aumentando a captação de água e nutrientes, e a sua acumulação nos tecidos vegetais (Duñabeitia et al. 2004; Fini et al. 2011).

Estudos evidenciaram que a inoculação de várias espécies de Quercus com fungos micorrízicos pode ter efeitos positivos no crescimento das plantas e na sua resposta fisiológica. As plantas micorrízicas demonstram uma maior aquisição de nutrientes, maior resistência a stresses ambientais, o que se traduz num aumento do crescimento (Núñez et al. 2006; Southworth et al. 2009; Sebastiana et al. 2013). Assim, a micorrização de plantas jovens de sobreiro em viveiros pode construir uma abordagem mais sustentável para aumentar o sucesso de novas plantações necessárias para acompanhar a crescente procura de cortiça (Sebastiana et al. 2013).

O uso de fungos micorrízicos em sistemas de produção vegetal constitui uma estratégia promissora para aumentar a produtividade da planta com reduzido impacto no meio ambiente, uma vez que o fungo atua como fertilizante, podendo reduzir-se os adubos sintéticos altamente prejudiciais para o meio ambiente (Sebastiana et al. 2016). Assim, informações mais detalhadas sobre os processos moleculares nas plantas hospedeiras das micorrizas são muito relevantes devido ao seu significado ecológico, à importância económica das espécies de árvores envolvidas e ao interesse em explorar essa simbiose para maximizar a produtividade e a sustentabilidade das árvores (Sebastiana et al. 2016). Estudos moleculares, incluindo investigações de perfil genético em larga escala, mostraram que as alterações morfológicas e fisiológicas associadas ao desenvolvimento das micorrizas são acompanhadas por alterações na expressão de genes e proteínas em ambos os parceiros da associação (Johansson et al. 2004; Flores Monterroso et al. 2013; Sebastiana et al. 2014, 2017).

1.2 Análise de dados de ómica

O acesso a conjuntos de dados de ómica em larga escala (genómica, transcriptómica, proteómica, metabolómica, metagenómica, fenómica, etc.) revolucionou a Biologia e levou ao surgimento de abordagens sistémicas para avançar a compreensão dos processos biológicos. (Misra et al. 2019)

Em experiências com dados de ómicas, tipicamente milhares de hipóteses são testadas simultaneamente, cada uma com base em muito poucas réplicas independentes. Testes tradicionais como o teste-t mostraram um fraco desempenho com este tipo de dados. Aqui, o baixo número de réplicas é frequente, chegando a ser de apenas duas ou três. Isto acontece quer pelo elevado custo de realização das experiências (fator mais importante), quer pelo uso de tecnologia muito sofisticada que requer muito tempo, ou até mesmo pelo tempo de recolha/crescimento das amostras biológicas.

1.3 Objetivos do estudo

O principal objetivo do trabalho foi efetuar a análise estatística de dados quantitativos de metabolómica gerados pela técnica de DI-FTICR-MS (*Direct Infusion Fourier Transform Ion Cyclotron Resonance Mass Spectrometry*), no sentido de identificar e quantificar os metabolitos que caracterizam a raíz de sobreiro quando esta se encontra em associação simbiótica com fungos micorrízicos. Os investigadores que trabalham com dados de metabolómica utilizam geralmente o *software* MetaboAnalyst (Chong, J. et al. 2018) para identificar os metabolitos que se expressam diferencialmente nas duas condições, experimental e controlo (plantas micorrizadas e não micorrizadas, neste caso). Este *software* inclui diferentes métodos estatísticos/matemáticos, mas consideramos que funciona como uma "caixa negra", no sentido em que não há uma indicação sobre a melhor opção em determinado contexto, nem são fornecidos detalhes sobre os métodos em si.

Contudo, existem hipóteses de escolha relativamente ao que se pretende fazer aos dados em várias etapas, tais como:

Estimação de *missing values* – Através de métodos como Análise de componentes principais (PCA), *k*-vizinho mais próximo (KNN), Análise de componentes principais probabilísticas (PPCA), Método de análise de componentes principais bayesianas (BPCA) e Decomposição em valores singulares (SVD).

Filtragem dos dados − Através de opções como Amplitude interquartil (IQR), Desvio padrão (SD), Desvio absoluto mediano (MAD), Desvio padrão relativo (RSD = SD/mean), Desvio padrão relativo não paramétrico (MAD/median), Valor médio da intensidade e Valor da intensidade mediana.

Normalização dos dados − tendo em conta a amostra biológica, os dados e a escala.

Por último, a escolha entre vários métodos estatísticos para identificação dos metabolitos acumulados, sendo o mais frequentemente utilizado o *Partial Least Squares - Discriminant Analysis* (PLS-DA), sendo possível a escolha entre *t-statistics*, SAM (*Significance Analysis of Microarrays*) e outros mais.

O presente trabalho, para além de dar resposta ao problema da análise de dados de metabolómica com recurso às ferramentas convencionais, tem como objetivo secundário apresentar alternativas através da exploração da possibilidade de aplicação de metodologias desenvolvidas no âmbito da comparação da expressão de genes em indivíduos sujeitos a condições distintas, tais como, os *packages* limma (*Linear Models for Microarray Data*) e RankProd (*Rank Products*), disponíveis no R através da plataforma Bioconductor (Breitling et al. 2004; Ritchie et al. 2015).

Contrariamente ao que acontece com os níveis de expressão dos genes, que compreendem valores reais positivos, os dados de metabolómica analisados neste projeto correspondem a valores inteiros positivos incluindo zero (contagens), implicando a utilização de metodologias estatísticas diferentes. No entanto, o *package* limma inclui uma adaptação para dados de sequenciação (contagens) e, por isso, acreditamos que poderá também ser aplicado a dados de metabolómica. Por outro lado, a aplicação do método *Rank Products* a dados de natureza contínua pode fazer sentido porque se baseia nas ordens dos valores observados. O facto de as intensidades assumirem uma grande variação de valores e o facto de haver uma normalização que praticamente exclui a existência de zeros, diminui a existência de um número elevado de empates e poderá proporcionar a obtenção de bons resultados com o *Rank Products*.

No capítulo seguinte (capítulo 2) é apresentada a descrição deste estudo. No terceiro capítulo encontra-se descrita a metodologia estatística considerada necessária à obtenção de resultados que permitam responder aos objetivos do estudo. No quarto capítulo são descritos os resultados estatísticos; e por fim, no quinto capítulo, apresenta-se a discussão e conclusões finais.

Capítulo 2

Descrição do Estudo

O procedimento laboratorial do presente estudo foi realizado na FCUL, tendo a obtenção das plantas micorrizadas e extração dos metabolitos sido efetuada no laboratório de Plant Functional Genomics do centro de investigação do Instituto de Biossistemas e Ciências Integrativas (BioISI), e a análise química dos compostos, pelo laboratório de FTICR e espectrometria de massa estrutural do centro de química e bioquímica.

Todos os dados analisados no presente projeto foram fornecidos por Mónica Sebastiana, investigadora pós-doutorada do Plant Functional Genomics Group (BIOISI-FCUL) que iniciou um projeto que tem como principal objetivo a caracterização do metaboloma da raíz micorrizada do sobreiro, ou seja, a identificação do conjunto total dos metabolitos (compostos biológicos) induzidos ou reprimidos quando a raíz do sobreiro se encontra micorrizada. A motivação para este estudo reside no facto de a micorrização trazer vantagens para a planta, incluindo maior crescimento e capacidade acrescida para resistir a stresses ambientais (e.g. seca, salinidade do solo) e doenças. Assim, a micorrização pode constituir uma solução natural alternativa à utilização de adubos químicos, fungicidas e pesticidas, que são altamente prejudiciais para o ambiente.

Para a caracterização do metaboloma foram comparadas raízes micorrizadas e não micorrizadas (controlos) de maneira a se poderem identificar os compostos diferencialmente acumulados (induzidos ou reprimidos nas raízes micorrizadas). Após extração dos metabolitos de ambos as amostras de raíz, recorreu-se à técnica de DIFTICR-MS (*Direct Infusion Fourier Transform Ion Cyclotron Resonance Mass Spectrometry*) que permite identificar e quantificar os metabolitos presentes numa mistura.

2.1 Protocolo Experimental

Nesta experiência foram plantadas sementes de sobreiro durante o mês de outubro do ano dois mil e dezoito. No ano seguinte, durante o mês de março, o fungo micorrízico foi inoculado nas plantas que, entretanto, já tinham germinado. (Fig. 2.1)

Figura 2.1 - Plantação de sobreiros micorrizados (Créditos: Mónica Sebastiana)

Posteriormente, dois a quatro meses após a inoculação, as raízes das plantas micorrizadas e controlos foram recolhidas para análise, possibilitando a identificação dos metabolitos. O estudo envolveu seis réplicas de plantas de sobreiros – sendo três micorrizadas e três utilizadas como grupo de controlo não micorrizado, para efeitos de comparação. Cada réplica consistia em um pool de raízes de três plantas. Os metabolitos de cada réplica foram extraídos utilizando quatro frações ou solventes: água (H2O), metanol (MeOH), acetonitrilo (ACN) e clorofórmio (orgânica), de acordo com um protocolo desenvolvido anteriormente e compatível com a técnica de FTICR (Maia et al. 2016). O extrato de cada fração foi analisado por FTICR MS em modo de ionização positivo (ESI+) e negativo (ESI-), de acordo com Maia et al. (2016). Para cada réplica foi obtido o conjunto de massas correspondentes aos diferentes metabolitos ou compostos biológicos (e.g. carboidratos, lípidos, ácidos nucleicos, compostos fitoquímicos, etc), apresentando cada uma delas uma intensidade que permite inferir a sua quantidade nessa réplica e identificar os compostos diferencialmente acumulados por comparação entre as réplicas micorrizadas e os controlos. Estes compostos são indicadores de possíveis informações de que essa massa está envolvida na simbiose entre o sobreiro e o fungo micorrízico, possibilitando novo conhecimento sobre as vias metabólicas envolvidas na micorrização.

2.2 Variáveis em Estudo

A variável de interesse neste estudo corresponde às intensidades, sendo esta medida através do método FTICR MS já referido. Trata-se de uma variável discreta com valores em $\mathbb{N}_0$.

A variável "Massa" (m.z) representa uma componente biológica associada a um determinado metabolito (carboidratos, lípidos, proteínas, etc).

Existem 6 réplicas biológicas para a variável intensidade: 3 sob a condição micorrizadas (Mic_1, Mic_2 e Mic_3) e 3 de controlo (Ctrl_1, Ctrl_2 e Ctrl_3), conforme se pode ver na tabela 2.1. Cada linha indicada na tabela corresponde a uma massa com as seis respetivas intensidades.

Quanto mais elevado for o valor da intensidade, mais quantidade dessa massa existe nessa dada réplica.

Tabela 2.1 – Exemplo dos dados obtidos, para 4 massas (1.ª coluna). Nas colunas Mic_ constam as intensidades de cada massa nas três amostras biológicas micorrizadas e nas colunas Ctrl_ estão as intensidades das três amostras biológicas de controlo.

Massas	Mic_1	Mic_2	Mic_3	Ctrl_1	Ctrl_2	Ctrl_3
227.20227	1332626	898984	738127	596974	785105	1089036
255.2335133	1471012	1449786	2193816	2439290	1831988	1733358
265.1083	944810	742478	548942	0	0	0
265.14839	1683460	1078075	2403660	2915765	2059001	3364050

2.3 Dimensões (Compostos Biológicos)

Os compostos biológicos são massas, as quais podem ser carboidratos, lípidos, ácidos nucleicos, compostos fitoquímicos, etc. São usadas oito frações que correspondem aos quatro solventes (água, metanol, acetonitrilo e clorofórmio) cada um associado a dois métodos de ionização (ESI+ e ESI-) relativamente à técnica de FTICR MS, conforme se pode ver na Tabela 2.2.

No total temos 6 amostras biológicas, já mencionadas anteriormente, para cada um dos conjuntos (solventes × métodos de ionização), perfazendo um total de 8 diferentes casos. Em termos de dimensão da amostra, esta varia de conjunto para conjunto, fazendo assim com que existam diferentes componentes biológicas associadas a tanto diferentes como iguais metabolitos.

No sentido de remover logo de início linhas que contenham um número bastante elevado de zeros, procede-se a uma primeira filtragem. Constatou-se, de imediato, que mais de metade das massas consideradas não acusavam quaisquer valores, ou seja, não existia qualquer quantidade dessa massa em nenhumas réplicas.

A segunda fase filtragem consiste na aplicação da função varFilter (integrada no *package* genefilter do R, Gentleman et al. 2019), um método extra de filtragem que tem como objetivo remover massas que exibam fraca variância das intensidades. De seguida, apresenta-se o número de massas iniciais (número inicial de linhas), o número de massas após a 1.ª filtragem e o número de massas após aplicação do varFilter, para as oito frações.

.

Tabela 2.2 – Tabela com os resultados das filtragens para cada um dos (Solvente x Métodos de Ionização), tendo em contas as massas obtidas inicialmente (Dados Iniciais), depois de uma 1ª Filtragem e Depois do varFilter.

Solvente x Métodos de Ionização	Dados Iniciais	Depois da 1ªFiltragem	Depois do varFilter
Fase Orgânica (clorofórmio) e Ionização por electrospray Negativo	104	35	31
Fase Orgânica (clorofórmio) e Ionização por electrospray Positivo	6296	768	691
ACN e Ionização por electrospray Negativo	729	194	174
ACN e Ionização por electrospray Positivo	5421	334	300
H2O e Ionização por electrospray Negativo	413	142	127
H2O e Ionização por electrospray Positivo	6879	770	693
MeOH e Ionização por electrospray Negativo	717	213	191
MeOH e Ionização por electrospray Positivo	6401	853	767

Tal como foi referido anteriormente, o processo de filtragem é muito importante porque se verificou que havia muitos zeros, indicando que algumas linhas (massas) não deveriam ser consideradas para o estudo.

Neste trabalho a primeira filtragem consistiu em 3 fases, como se descreve de seguida.

Fases das Filtragens

*** 1.ª Fase:**

Eliminar todas as linhas cujas réplicas micorrizadas ou de controlo, apresentem todos os índices nulos. Quando os níveis de intensidade são nulos nas amostras consideradas, como é possível constatar na Tabela 2.2 (linha 3), isto significa que não existe qualquer quantidade dessa massa nessa réplica.

* **2.ª Fase:**

Quando em três réplicas, quer micorrizadas ou controlo, duas destas não apresentem qualquer índice de intensidade, assume-se como nula a única que apresenta intensidade.

* **3.ª Fase**

Quando, se apresentam pelo menos dois índices de intensidade, pertencentes quer a elementos de controlo ou micorrizadas, considerou-se necessária a sua inclusão no estudo, uma vez que, esses valores poderiam ser importantes para a identificação das massas diferencialmente expressas. O valor nulo será substituído pela média das duas intensidades observadas no mesmo grupo (Mic ou Ctrl).

varFilter

Recorreu-se ao método varFilter para possibilitar uma filtragem extra na tentativa de excluir algumas massas que exibam fraca variância de intensidades das seis réplicas. Este consiste num filtro biológico e num filtro de variação estatística que visa remover contaminantes óbvios nos dados de espectrometria de massa de purificação por afinidade (Bourgon et al. 2010).

A remoção não deverá representar um número muito elevado de massas. Esta função está incluída no *package* genefilter (Gentleman et al. 2019).

Na primeira filtragem foi onde se observou uma maior quantidade de linhas a serem eliminadas, isto aconteceu pelos dados em causa na sua maioria terem uma grande quantidade de valores nulos e não poderem ser considerados para a amostra. Na filtragem do varFilter, este removeu linhas que exibissem uma variância reduzida entre réplicas como já dito anteriormente, constatando-se uma quantidade pequena destes casos.

Capítulo 3

Metodologia Estatística

Neste capítulo serão abordados todos os métodos estatísticos que foram utilizados no decorrer do estudo. Será dada uma visão teórica por detrás das variadas metodologias que foram aplicadas no sentido de se responder às questões indicadas nos objetivos.

Os métodos paramétricos requerem certos pressupostos (e.g. normalidade, igualdade de variâncias), e é com frequência que esta tipologia de teste se orienta pela premissa de que as referidas condições sejam apuradas. No entanto, os métodos não paramétricos exigem suposições mínimas sobre a forma distributiva da população. É uma opção útil quando corretamente utilizada, em que a pré-condição de normalidade não existe. No presente projeto iremos abordar os dois tipos de métodos (paramétricos e não paramétricos), para de seguida comparar os resultados obtidos.

Tal como foi previamente indicado na descrição do estudo, antes de se proceder à análise estatística, é necessário filtrar os dados. A primeira filtragem baseou-se na eliminação de linhas de acordo com regras associadas às médias das intensidades e à quantidade destas que se consideraram inexistentes entre os dois grupos (Controlo vs Micorrizadas). De seguida, recorreu--se a um método auxiliar de filtragem (varFilter) para que fosse possível obter uma amostra mais representativa, eliminando-se assim algumas massas que exibiam pouca ou quase nenhuma variação entre as amostras biológicas (réplicas).

3.1 Filtragem baseada na variabilidade

O método aqui intitulado de "varFilter" (variation-based filtering) foi aplicado numa fase inicial da análise dos dados, depois de uma filtragem usada conforme três fases enunciadas já anteriormente. Apesar deste método ser mais usualmente aplicado em dados de microarrays, recorreu-se a este como uma solução para remover linhas que exibissem uma variância reduzida entre réplicas. Segundo Bourgon et al. (2010), esta filtragem será vantajosa pois irá remover massas que exibam pouca variação ou um sinal consistentemente baixo entre amostras. Assim, é dado uso à função varFilter na eliminação destas massas, que segundo Bourgon et al. (2010) permite controlar o erro tipo I.

Como se constatou neste estudo, o número de massas inicial em comparação com o número de massas após a filtragem (primeira e segunda filtragem), foi reduzido para mais de metade – de acordo com Bourgon et al. (2010) uma filtragem enriquecedora será aquela que de

facto irá remover linhas que não trazem qualquer informação, e possibilitar assim uma seleção mais precisa sobre os compostos diferencialmente acumulados.

Recorrendo à função varFilter do R, este filtro será calculado através IQR (Amplitude Interquartil), dada pela expressão:

$$IQR = Q_3 - Q_1 \tag{3.1}$$

A amplitude interquartil resume-se à distância entre o primeiro quartil (Q_1) e o terceiro quartil (Q_3) em que Q_1 é dado pela mediana dos dados à esquerda da mediana (Q_2) e Q_3 pela mediana dos dados à direita da mediana. Esta opção de escolha, por definição, é motivada pela observação de que os genes não expressos são detetados com maior confiança desse houver pouca variabilidade entre as amostras, e também por se ter a informação de que IQR é robusto a *outliers*. (Bourgon et al. 2010).

Por definição, a função varFilter assume o argumento var.func = "IQR" (Amplitude Interquartil), existindo outras opções de escolha.

3.2 Método Rank Products

O método *Rank Products* (*package* RankProd do Bioconductor) foi desenvolvido com a intenção de possibilitar a deteção de genes diferencialmente expressos em *microarrays* de dois canais (dados emparelhados) sob duas condições experimentais (Breitling et al. 2004). O *Rank Products* é um método não paramétrico baseado em ordens, ou seja, baseia-se nas ordens dos níveis de expressão para cada réplica, calculando o produto dessas ordens para cada gene. Um teste de permutação serve de base para a identificação de genes que se expressam de forma significativamente diferente sob duas condições. Em 2010, Koziol, J. adaptou este método para o caso de duas amostras independentes.

Este método provém do raciocínio biológico que identifica genes que são na maioria altamente expressos em várias réplicas sob uma condição experimental e menos expressos noutra condição (Hong et al. 2006), tratando-se de um método que aparenta ser robusto, com maior sensibilidade e especificidade que o teste-t na presença de hipóteses múltiplas e poucas réplicas. (Koziol, J. 2010).

Segundo Koziol (2010), o método *Rank Products* aplica-se em estudos que originam dados que consistem em *n* compostos (originalmente genes, mas podendo ser alargados a metabolitos, proteínas, etc.) com níveis de expressão de *k* amostras biológicas replicadas sob duas condições experimentais diferentes (mutação *vs*. controlo, por exemplo).

As hipóteses a testar tendo em conta este estudo são:

H_0: Os genes são não diferencialmente expressos

vs

H_1: Existe pelo menos um gene diferencialmente expresso

O nível de expressão do g-ésimo gene na j-ésima réplica biológica da m-ésima condição experimental é dado por X_{gjm}, onde $g = 1, \dots, n,\ j = 1, \dots, k_m \quad e \quad m = 1, 2$. De seguida, são ordenados os níveis de expressão

$X_{1jm}, X_{2jm}, \dots, X_{njm}$ dentro de cada réplica biológica j, onde

$$R_{gjm} = rank_g(X_{gjm}), \qquad g = 1, \dots, n \tag{3.2}$$

Uma versão simplificada para o caso de duas amostras independentes (condições experimentais) do método *Rank Products*, segundo Breitling et al. (2004), para o g-ésimo gene é dada por:

$$RP_g = \left(\prod_{j=1}^{k_1} R_{gj1} \right)^{\frac{1}{k_1}} \div \left(\prod_{j=1}^{k_2} R_{gj2} \right)^{\frac{1}{k_2}} \tag{3.3}$$

Tem-se que RP_g é a média geométrica das ordens para o g-ésimo gene da amostra 1, sendo de seguida dividida pela média geométrica das ordens para g-ésimo gene da amostra 2.

A necessidade de recorrer a um grande número de permutações e tempo de computação para calcular os valores-p para testar a expressão diferencial em cada gene (Eisinga et al. 2013), levou Koziol (2010) a propor uma estatística de teste com transformação logarítmica considerando a distribuição gama para aproximar os valores-p.

Uma formulação alternativa envolve a estatística RP_g mas usando a transformação logarítimica, garantindo que os níveis de significância para RP_g e $log(RP_g)$ são idênticos:

$$log(RP_g) = \left(\frac{1}{k_1}\right) \times \sum_{j=1}^{k_1} log(R_{gj1}) - \left(\frac{1}{k_2}\right) \times \sum_{j=1}^{k_2} log(R_{gj2}) \qquad (3.4)$$

Vantagens do método Rank Products

Uma das vantagens é o bom desempenho em amostras de pequenas dimensões com milhares de hipóteses em simultâneo e, outra vantagem, é a facilidade com que os resultados de várias plataformas experimentais podem ser combinados em uma análise (Breitling et al. 2004). Também pode ser aplicado a estudos de proteoma e metaboloma, onde listas de expressão de proteínas ou metabolitos alterados são produzidas por géis 2D ou espectrometria de massa (Breitling et al. 2004).

Razão das falsas descobertas (FDR)

De acordo com Reiner et al. (2003), a razão das falsas descobertas (FDR - *False Discovery Rate)* é definida da seguinte maneira:

Considere-se uma família de n hipóteses nulas testadas simultaneamente em que $m0<m$ são verdadeiras. Para cada hipótese H_{0i} (i=1,…,n), o valor observado de uma estatística de teste é calculado produzindo um valor-p correspondente: p_i. R denota o número de hipóteses rejeitadas e V o número de hipóteses nulas verdadeiras rejeitadas (erradamente). Tem-se que $Q = \frac{V}{R}$, quando $R > 0$ e $Q = 0$, caso $R = 0$. Como não é possível calcular o número de falsos positivos (V), considera-se o número esperado de falsos positivos, $E(V)$. Então, a FDR é definida como

$$FDR = E\,(Q) = E(V)/R \qquad (3.5)$$

Quanto menor for o valor de FDR melhor, pois significa que existem poucos falsos positivos entre os selecionados como positivos (hipóteses nulas rejeitadas). De acordo com a lituratura, valores aceitáveis para a FDR variam até um máximo de 0.2.

Como alternativa ao método *Rank Products*, recorreu-se ao *package* limma do Bioconductor, que tem por base modelos lineares para analisar a expressão diferencial de dados de *microarrays* e dados de RNA-Seq (dados de sequenciação). Tal como no caso de dados de RNA-Seq, os dados aqui analisados são dados de contagens e, como tal, ter-se-á ainda que recorrer à função voom, do limma, especificamente desenvolvida para contemplar a análise de dados de sequenciação.

3.3 Método "voom-limma"

Função voom

O voom foi desenvolvido para dados de sequenciação e o procedimento é feito de acordo com Law et al. (2014), onde os dados (contagens de *reads*) começam por ser transformados tendo por base a aplicação do logaritmo de base 2. Os dados analisados neste trabalho, também correspondem a contagens. A normalização opcional é executada usando *normalizeBetweenArrays*.

Após a aplicação da função voom, são utilizadas as seguintes funções do *package* limma: LmFit, como o cálculo da estatística t-moderada; e Ebayes; que permitem identificar genes com expressão diferencial. (Law et al. 2014).

Package limma

O *package* limma, tal como foi referido anteriormente, tem como objetivo encontrar genes diferencialmente expressos através da análise de expressão desses mesmos genes, obtida através de técnicas experimentais como *microarrays* ou sequenciação (RNA-Seq).

Numa análise recorrendo a este *package* e de acordo com Ritchie et al. (2015), este *package* integra um número de princípios estatísticos de um modo a que seja eficaz para estudos de expressão em larga escala, organizados numa matriz de valores, em que em cada linha corresponde a um gene e cada coluna a uma réplica biológica. Os princípios estatísticos envolvem, por um lado, o ajustamento de um modelo linear a cada linha de dados tirando proveito da flexibilidade desses modelos de várias maneiras, como por exemplo, lidar com hipóteses biológicas. Por outro lado, aproveita os dados genómicos para emprestar força entre os modelos em termos genéticos, permitindo diferentes níveis de variabilidade entre genes e entre amostras, e tornando as conclusões estatísticas de maior confiança quando o número de amostras é pequeno. Todas as características dos modelos estatísticos podem ser utilizadas não apenas para análises de expressão em termos genéticos, mas também para análises de nível superior de assinaturas de expressão genética.

Para cada gene g (g=1,…,n), temos um vetor aleatório expressão genética ($\boldsymbol{Y_g}$ = (Y_1,…,Y_k)), cujas componentes seguem um modelo normal, e uma matriz de desenho X (*design matrix* - matriz de valores de variáveis explicativas do conjuntos de dados) que relaciona estes valores com um vetor de coeficientes de interesse ($\boldsymbol{\alpha_g} = (\alpha_{g1}, \alpha_{g2})$). Nesta matriz X, temos representada em cada linha uma massa e nas colunas temos 3 que são de amostras biológicas micorrizadas e 3 que são de amostras biológicas de controlo, esta matriz indica as comparações que são feitas.

De acordo com os modelos lineares tem-se que:

$$E(\boldsymbol{Y_g}) = \boldsymbol{X\alpha_g} \tag{3.6}$$

e

$$Var(\boldsymbol{Y}_g) = W_g \times \sigma_g^2 \tag{3.7}$$

Também se salienta que, $\boldsymbol{\alpha_g}$ e $\sigma_g{}^2$ são parâmetros específicos para cada gene , W_g é uma matriz de pesos definida não negativa conhecida e a hipótese nula de interesse é $\beta_g = 0$ contra a hipótese alternativa $\beta_g \neq 0$, sendo β_g dado pelo contraste $\alpha_{g1} - \alpha_{g2}$. A matriz *X* é dada por

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \tag{3.8}$$

tendo em conta os nossos dados.

Este *package* inclui métodos que permitem o tratamento de informação recorrendo à estatística t-moderada ou ao método empírico de Bayes.

As suposições distributivas podem ser resumidas por :

$$\hat{\beta}_g | \beta_g, \sigma_g^2 \sim N(\beta_g, \vartheta_g \times \sigma_g^2) \tag{3.9}$$

e

$$s_g^2 | \sigma_g^2 \sim \frac{\sigma_g^2}{d_g} \times \chi_{dg}^2 \tag{3.10}$$

onde d_g são os graus de liberdade do erro para o modelo linear para o gene g e ϑ_g é o j-ésimo elemento da diagonal de $C^T V_g C$, em que C é o vetor de contrastes e $C^T = [1 \ \ -1]$. Sob estes pressupostos, a estatística-t

$$t_g = \frac{\widehat{\beta}_g}{s_g * \sqrt{\vartheta_g}} \tag{3.11}$$

segue uma distribuição-t aproximada com d_g graus de liberdade.

Estatística t-moderada

Segundo Smyth (2004), dado o grande número de modelos lineares em termos genéticos vindos de experiências com *microarrays*, existe uma necessidade de tirar vantagem da estrutura paralela em que o mesmo modelo é ajustado em cada gene. De seguida, vai ser descrito um modelo hierárquico simples para descrever este efeito paralelo, tendo como principal foco descrever como os coeficientes β_g, desconhecidos, e variâncias $\sigma_g{}^2$, desconhecidas, variam entre genes assumindo distribuições *a priori* para estes parâmetros (Smyth, 2004).

Smyth (2004), assume uma distribuição *a priori* de $\frac{1}{\sigma_g^2}$ da seguinte forma:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 * s_0^2} \times \chi^2_{d_0} \tag{3.12}$$

Para cada gene g, assume-se que β_g diferente de zero com probabilidade:

$$P(\beta_g \neq 0\ ) = p \tag{3.13}$$

em que p é a proporção esperada dos genes verdadeiramente diferencialmente expressos. Para $\beta_g \neq 0$, assume-se *a priori* que a sua variância é ϑ_0 , pelo que,

$$\beta_g | \sigma_g^2, \beta_g \neq 0 \sim N(0, \vartheta_0 \sigma_g^2) \tag{3.14}$$

descreve a distribuição do *log-fold change* (β_g: logaritmo da razão da quantidade de nutrientes acumulados nas micorrizadas e de controlo) para genes que sejam diferencialmente expressos.

Temos segundo o modelo hierárquico, a média *a posteriori* de σ_g^{-2} dado s_g^2 dada por:

$$\tilde{s}_g^2 = \frac{d_0 \times s_0^2 + d_g \times s_g^2}{d_0 \times d_g} \tag{3.15}$$

A estatística t-moderada é definida por

$$\tilde{t}_g = \frac{\tilde{\beta}_g}{\tilde{s}_g \times \sqrt{\vartheta_g}} \tag{3.16}$$

A estatística t-moderada segue aproximadamente uma distribuição-t tendo em conta a hipótese nula H_0: $\beta_g = 0$ com $d_g + d_0$ graus de liberdade. Os graus de liberdade adicionados para $t_{\tilde{g}}$ sobreposto a t_g reflete a informação extra, com base no modelo hierárquico, a partir do conjunto de genes para a inferência sobre cada gene individual. É de constatar que este resultado vindo desta distribuição assume d_0 e s_0^2 como valores dados. Estes valores são estimados de acordo com o procedimento descrito por Smyth (2004) (secção 6).

Estatística B – logaritmo das possibilidades *a posteriori*

Segundo Lönnstedt & Speed (2001), os dados referentes a todos os genes num conjunto replicado de experiências são combinados em estimativas de parâmetros que por sua vez também seguem uma distribuição de probabilidade, de acordo com a metodologia bayesiana. Essas estimativas de parâmetros são então combinadas no nível do gene para formar uma estatística B, que corresponde ao logaritmo das possibilidades (*log odds*) *a posteriori*.

De acordo com Smyth (2004), as distribuições marginais da estatística t-moderada $t_{\tilde{g}}$ e de s_g^2, facilitam o cálculo das possibilidades *a posteriori* (O_g) de cada gene *g* ser diferencialmente expresso, tendo por base β_g. Note-se que, tal como Smyth (2004) indica, $\tilde{t}$ e s^2 são independentes com

$$s^2 \sim s_0^2 \times F_{(d,d_0)} \tag{3.17}$$

e

$$\tilde{t}_g \mid \beta_g = 0 \sim t_{(d_0+d)} \tag{3.18}$$

A derivação acima passa pelo mesmo com $\beta_g \neq 0$, a única diferença é que

$$\tilde{t}_g \mid \beta_g \neq 0 \sim \left(1 + \frac{\vartheta_0}{\vartheta_g}\right)^{\frac{1}{2}} \times t_{(d_0+d)} \tag{3.19}$$

A distribuição marginal de $\tilde{t}$ sobre todos os genes é, portanto, uma mistura de uma distribuição-t em escala e uma distribuição-t normal com proporções de mistura p e $1-p$ respetivamente.

A possibilidade de o g-ésimo gene ser diferencialmente expresso (β_g diferente de zero) *vs.* não ser diferencialmente expresso, condicional aos dados, é dada por

$$O_g = \frac{p(\beta_g \neq 0 \mid \tilde{t}_g, s_g^2)}{p(\beta_g = 0 \mid \tilde{t}_g, s_g^2)} = \frac{p(\beta_g \neq 0, \tilde{t}_g, s_g^2)}{p(\beta_g = 0, \tilde{t}_g, s_g^2)} = \frac{p}{1-p} \times \frac{p(\tilde{t}_g \mid \beta_g \neq 0)}{p(\tilde{t}_g \mid \beta_g = 0)} \tag{3.20}$$

desde que $\tilde{t}_g$ e s_g^2 sejam independentes e a distribuição de s_g^2 não depende de β_g.

Tendo em conta a densidade de $t_{\tilde{g}}$ apresentada anteriormente, a expressão (3.20) vem

$$O_g = \frac{p}{1-p} * \left(\frac{\vartheta_g}{\vartheta_g + \vartheta_0}\right)^{\frac{1}{2}} \times \left(\frac{\tilde{t}_g + d_0 + d_g}{\tilde{t}_{gj} * \frac{\vartheta_g}{\vartheta_g + \vartheta_0} + d_0 + d_g}\right)^{\frac{1+d_0+d_g}{2}} \tag{3.21}$$

Esta fórmula está em concordância com Lönnstedt & Speed (2002). Se $d_0 + d_g$ for grande, as possibilidades *a posteriori* reduzem para

$$O_g = \frac{p}{1-p} \times \left(\frac{\vartheta_g}{\vartheta_g + \vartheta_0}\right)^{\frac{1}{2}} \times exp\left(\frac{\tilde{t}_g}{2} * \frac{\vartheta_{gj}}{\vartheta_g + \vartheta_0}\right) \tag{3.22}$$

A estatística B corresponde ao logaritmo das possibilidades *a posteriori*, ou seja, $B_g = log(O_g)$

De acordo com Lönnstedt & Speed (2002), a estatística está numa escala aceitável e é útil para ordenar genes e classificá-los como sendo diferencialmente expressos, ou não. Para a seleção dos genes diferencialmente expressos consideram-se aqueles que tenham valores $B_g > 0$. Isto implica que em (3.18) o numerador (equivalente à probabilidade de ser diferencialmente expresso) seja superior ao denominador (equivalente à probabilidade de não ser diferencialmente expresso).

3.4 Método PLS-DA

PLS-DA (*Partial Least Squares – Discriminant Analysis*) é uma técnica quimiométrica usada para otimizar a separação entre diferentes grupos de amostras, que é conseguido unindo duas matrizes de dados *X* (dados em bruto) e *Y* (grupos) (Gromski et al. 2015). Esta abordagem visa maximizar a covariância entre as variáveis independentes $X_1,\ldots,X_n$ (leituras; isto é, os dados metabolómicos) e a variável dependente correspondente à variável *Y* (classes; grupos) de dados multidimensionais, encontrando um subespaço linear de variáveis explicativas (Gromski et al. 2015).

A principal vantagem da abordagem PLS-DA é o manuseamento de dados altamente colineares e ruidosos, que são resultados muito comuns de experiências de metabolómica (Want & Masson, 2011). Além disso, fornece várias estatísticas, como os *loading weight*, VIP (*Variable Importance in Projection*) e coeficiente de regressão que podem ser usados para identificar as variáveis mais importantes (Mehmood et al. 2012; Mehmood et al. 2011; Krishnan et al. 2011)

Um requisito fundamental para que o PLS produza respostas significativas é alguma seleção preliminar de variáveis (Pérez-Enciso et al. 2003). Isto é feito através do VIP para cada variável, que é uma medida que estima a importância de cada variável na projeção usada num modelo de PLS. Este método é bastante popular na literatura do PLS e é definido como:

$$VIP_j = \left\{ \frac{p \times \sum_{h=1}^{m} \sum_k R^2(y_k, t_h) \times w_{hj}^2}{\sum_{h=1}^{m} \sum_k R^2(y_k, t_h)} \right\}^{\frac{1}{2}} \quad (3.23)$$

(Eriksson et al. 1999; Tenenhaus, 1998) para cada *j*-ésima variável preditora (independente) *j*=1, *p*, em que $R^2(a, b)$ representa a correlação ao quadrado entre *a* e *b*, e $t_h = X_{h-1} w_h$, em que X_{h-1} é a matriz residual na regressão de *X* nos componentes $t_1, \ldots, t_{h-1}$ e w_h é um vetor de norma 1 (no algoritmo de regressão PLS, t_h é construído com essa restrição de normalização). Observa-se que $w_j{}^2{}_h$ mede a contribuição de cada variável *j* para o *h*-ésimo componente do PLS (Pérez-Enciso et al. 2003).

Uso do MetaboAnalyst para dados de metabolómica

O MetaboAnalyst é frequentemente usado para a análise de dados de estudos de metabolómica. Este destina-se a ajudar os utilizadores a realizar análises de dados, visualizações e interpretação funcional em dados de metabolómica (Chong et al. 2018). O MetaboAnalyst oferece várias opções para processamento de dados metabolómicos, tais como: análise estatística multivariada, identificação de metabolitos, normalização, mapeamento de vias e representações gráficas. (Xia et al. 2009) Em particular, este suporta técnicas como: análises de FC (*FoldChange*), clustering hierarquico, PCA, PLS-DA, testes-t e vários outros métodos sofisticados de estatística e *machine learning.* (Xia et al. 2009)

Em particular, o MetaboAnalyst é capaz de processar uma ampla variedade de tipos de dados metabolómicos, incluindo tabelas de concentração de compostos (para metabolómica quantitativa), listas de pico de RMN / MS e espectros de GC / LC-MS (NetCDF, mzXML, mzDATA— para metabolómica quimiométrica). (Xia et al. 2009).

Capítulo 4

Resultados

Neste capítulo serão apresentados os resultados obtidos após a aplicação dos métodos *Rank Products*, "voom-limma" e PLS-DA (MetaboAnalyst) para identificação dos metabolitos diferencialmente acumulados nas raízes micorrizadas de sobreiro.

4.1 Método Rank Products

Aplicou-se o método *Rank Products* para as 8 frações, considerando-se 100 permutações - número preestabelecido pela função do R e que decidimos manter (as tabelas encontram-se em anexo). Os termos Up e Down que irão surgir de seguida estão associados às massas induzidas e reprimidas, respetivamente. Para cada uma das 8 frações, o nível de expressão do g-ésimo gene na j-ésima réplica biológica da m-ésima condição experimental é dado por X_{gjm}, onde neste caso, g corresponde ao nº de genes presentes em cada fração de dados, $j = 6$ réplicas biológicas e m = 2 tratamentos associados a Micorrizado e Controlo.

Os gráficos produzidos pelo *package* RankProd mostram o aumento no número de massas identificadas como diferencialmente expressas corresponde ao aumento da FDR. Para comparação entre ESI (positivo e negativo), para os 4 compostos são indicados com a cor vermelho todas as massas selecionadas como diferencialmente acumuladas. Note-se que o RankProd foi desenvolvido para dados de *microarrays*, pelo que no título dos gráficos surge o termo "genes" em vez de "massas". Vai ser considerada uma FDR de 0.2, valor que está dentro dos limites aceitáveis para esta razão, de forma a podermos encontrar um maior número de massas que serão depois avaliadas por uma equipa de biólogos no laboratório de FTICR e espectrometria de massa estrutural do centro de química e bioquímica

Numa análise comparativa entre ACN positivo (Up e Down) e negativo (Up e Down), o total de compostos diferencialmente acumulados para uma FDR inferior a 0.2, para o primeiro foi de, respetivamente, 15 para ambos, e para o segundo foi de, respetivamente, 12 e 10. Assim, verificou-se uma quantidade de compostos diferencialmente acumulados praticamente análoga tanto para ACN positivo como negativo.

De seguida, encontram-se os gráficos de identificação de massas reprimidas e induzidas com FDR<0.2, em que no eixo das abcissas temos o número de genes identificados, em que os *estimated pfp* são na realidade as FDR, e no eixo das ordenadas temos a estimativa da proporção de falsos positivos entre as hipóteses nulas rejeitadas. Estes gráficos são obtidos através da função topGene do *package* RankProd, que identifica genes diferencialmente expressos recorrendo ao método *Rank Product*s.,

ACN Positivo:

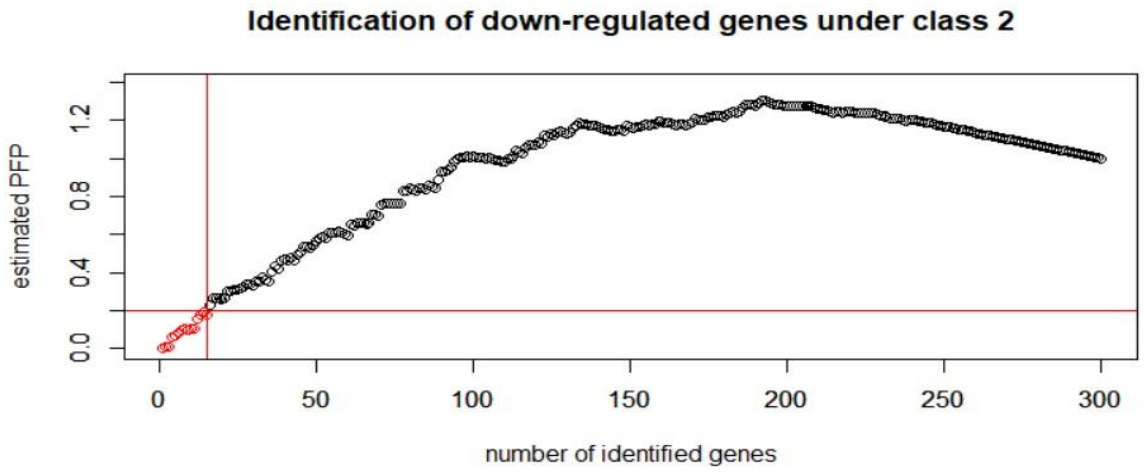

Figura 4.1 - Gráfico de identificação de massas reprimidas com FDR<0.2 para ACN Positivo

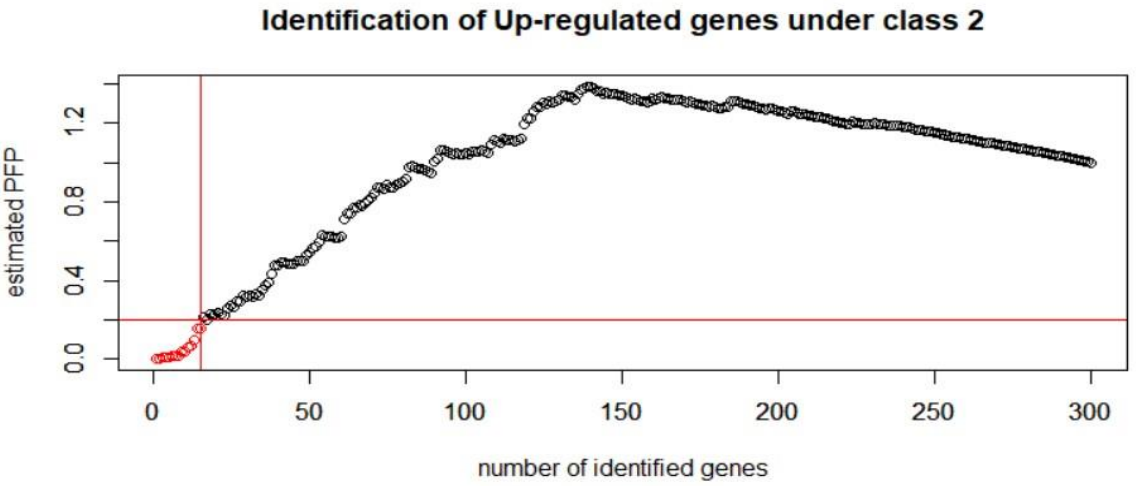

Figura 4.2 - Gráfico de identificação de massas induzidas com FDR<0.2 para ACN Positivo

ACN Negativo:

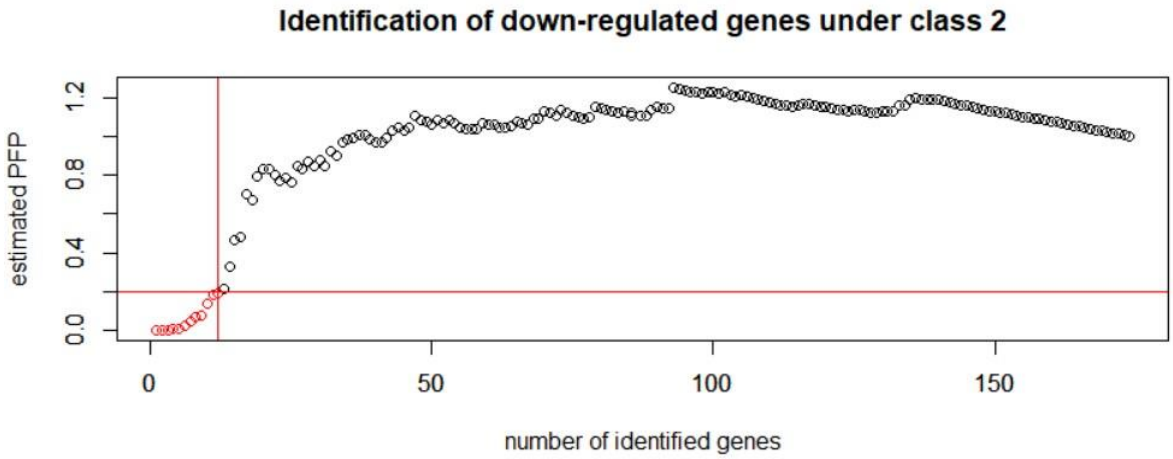

Figura 4.3 - Gráfico de identificação de massas reprimidas com FDR<0.2 para ACN Negativo

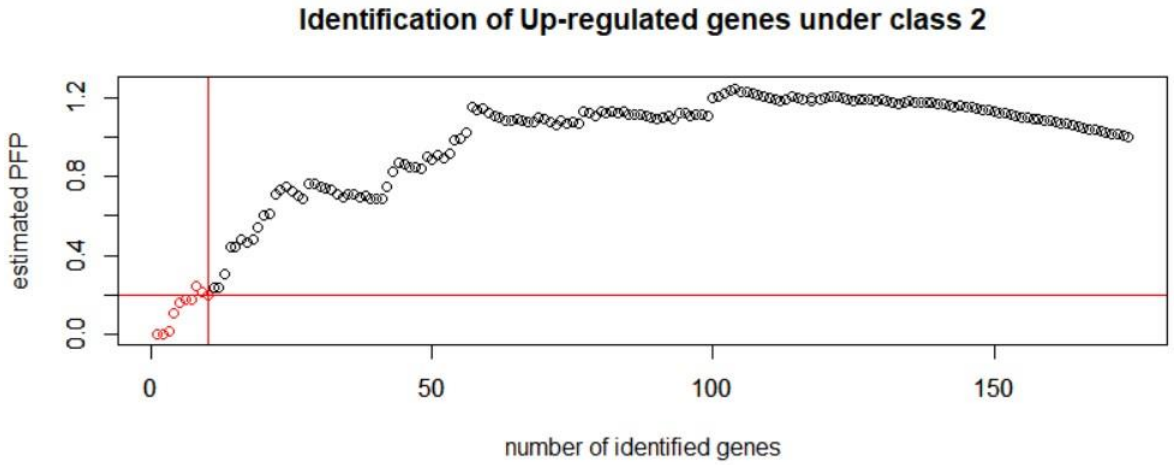

Figura 4.4 - Gráfico de identificação de massas induzidas com FDR<0.2 para ACN Negativo

Para as análises de H2O, MeOH e Fase Orgânica, estes foram os resultados obtidos:

Numa análise comparativa entre H2O positivo (Up e Down) e negativo (Up e Down), e um ponto de corte para uma FDR inferior a 0.2 (disponível em anexo), o total de compostos diferencialmente acumulados para o primeiro foi de, respetivamente, 19 e 39, e para o segundo foi de, respetivamente, 9 e 10. Assim, verificou-se uma quantidade de compostos diferencialmente acumulados mais acentuado para H2O positivo do que para o negativo.

Numa análise comparativa entre MeOH positivo (Up e Down) e negativo (Up e Down), o total de compostos diferencialmente acumulados para o primeiro foi de, respetivamente, 45 e 47, e para o segundo foi de, respetivamente, 14 e 16 para uma FDR inferior a 0.2 (disponível em anexo). Notou-se assim uma quantidade de compostos diferencialmente acumulados mais acentuada para o MeOH positivo comparativamente ao negativo. É de salientar que há muito mais compostos para o (ESI+) em comparação com o (ESI-).

Numa análise comparativa entre Orgânico positivo (Up e Down) e negativo (Up e Down), o total de compostos diferencialmente acumulados para o primeiro foi de, respetivamente, 49 e 46, e para o segundo foi de, respetivamente, 3 e 2, considerando uma FDR inferior a 0.2 (disponível em anexo). Verificando-se de novo assim uma quantidade de compostos diferencialmente acumulados mais acentuada para a fase orgânica na seleção positivo relativamente ao negativo, que era de esperar já que o número de massas é bastante inferior em Orgânico negativo.

Comentário global de todas as frações em conjunto

Na generalidade, pôde-se constatar que os compostos diferencialmente acumulados foram mais acentuados no caso positivo ao invés do negativo.

MeOH e a fase Orgânica, na totalidade, acusaram mais compostos diferencialmente acumulados, sendo que Orgânico negativo acusou menos compostos (total de 5), e Orgânico positivo acusou mais (total de 95).

Considerando uma FDR inferior a 0.2 foram selecionadas no total 351 massas, no entanto, a título de curiosidade e também para comparação com as massas selecionadas com um *cutoff* de 0.2, apresentamos o total de massas para um *cutoff* de 0.1:

MeOH:

- Para MeOH positivo $\begin{cases} 23 \text{ massas induzidas} \\ 30 \text{ massas reprimidas} \end{cases}$
- Para MeOH negativo: $\begin{cases} 13 \text{ massas induzidas} \\ 11 \text{ massas reprimidas} \end{cases}$

Fase Orgânica:

- Para Orgânico positivo $\begin{cases} 24 \text{ massas induzidas} \\ 31 \text{ massas reprimidas} \end{cases}$
- Para Orgânico negativo $\begin{cases} 2 \text{ massas induzidas} \\ 1 \text{ massa reprimida} \end{cases}$

ACN:

- Para ACN positivo $\begin{cases} 8 \text{ massas induzidas} \\ 13 \text{ massas reprimidas} \end{cases}$
- Para ACN negativo $\begin{cases} 9 \text{ massas induzidas} \\ 3 \text{ massas reprimidas} \end{cases}$

H2O:

- Para H2O positivo $\begin{cases} 13 \text{ massas induzidas} \\ 23 \text{ massas reprimidas} \end{cases}$
- Para H2O negativo $\begin{cases} 5 \textit{ massas induzidas} \\ 7 \textit{ massas reprimidas} \end{cases}$

Em termos biológicos, os resultados obtidos através deste método permitiram a identificação de:

- 275 compostos dif. exp. em ESI(+)
- 76 compostos dif exp. em ESI(-)

Perfazendo assim um total de 351 compostos diferencialmente acumulados, em que destes, 25 correspondem a compostos conhecidos nas bases de dados (MassTrix), ou seja, foi permitida a sua identificação (anotação).

Os metabolitos diferencialmente acumulados (Fig.4.5) pertencem a 4 classes de compostos de origem biológica: lípidos, compostos fitoquímicos, açúcares e amino ácidos. As classes com maior número de metabolitos diferencialmente acumulados foram as dos lípidos e

dos compostos fitoquímicos, o que indica que estas duas classes de compostos são as que mais se alteraram na raíz de sobreiro após a simbiose com o fungo micorrízico.

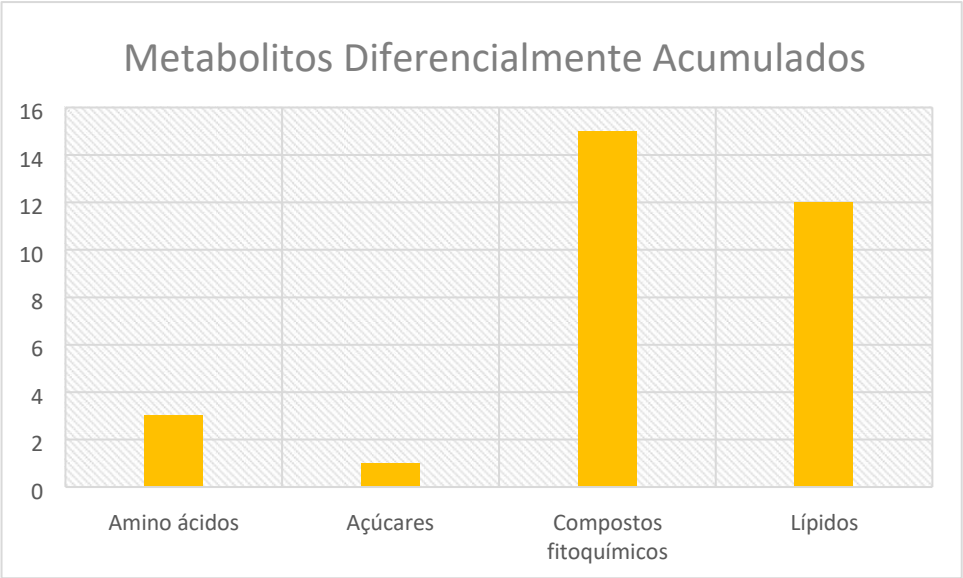

Figura 4.5 – Classes de compostos biológicos associados aos metabolitos diferencialmente acumulados identificados na raíz de sobreiro após micorrização com o fungo Pisolithus tinctorius

Dentro da classe dos compostos fitoquímicos, constituída por compostos do metabolismo secundário das plantas, foram identificados metabolitos pertencentes às classes dos alcaloides, flavonoides, terpenoides e fenilpropanoides (Fig. 4.6).

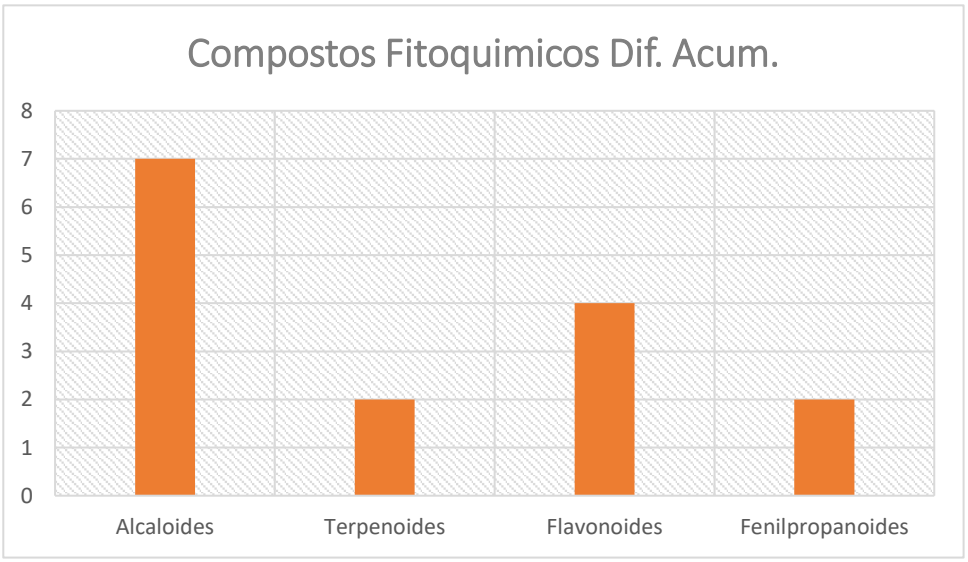

Figura 4.6 – Classes de compostos fitoquímicos associados aos metabolitos diferencialmente acumulados identificados na raíz de sobreiro após micorrização com o fungo Pisolithus tinctorius

No que respeita à acumulação diferencial dos metabolitos identificados pelo Rank Product, 10 foram induzidos e 14 reprimidos após micorrização da raíz de sobreiro. Este resultado mostra

que a micorrização parece afetar de forma positiva e negativa o mesmo número de metabolitos, o que está de acordo com trabalhos anteriores neste sistema biológico que mostram a mesma tendência para genes e proteínas (Sebastiana et al. 2014; 2017).

No grupo de metabolitos ativados, os mais representativos correspondem a compostos fitoquímicos pertencentes às classes dos flavonoides, terpenoides e alcaloides (Fig. 4.7).

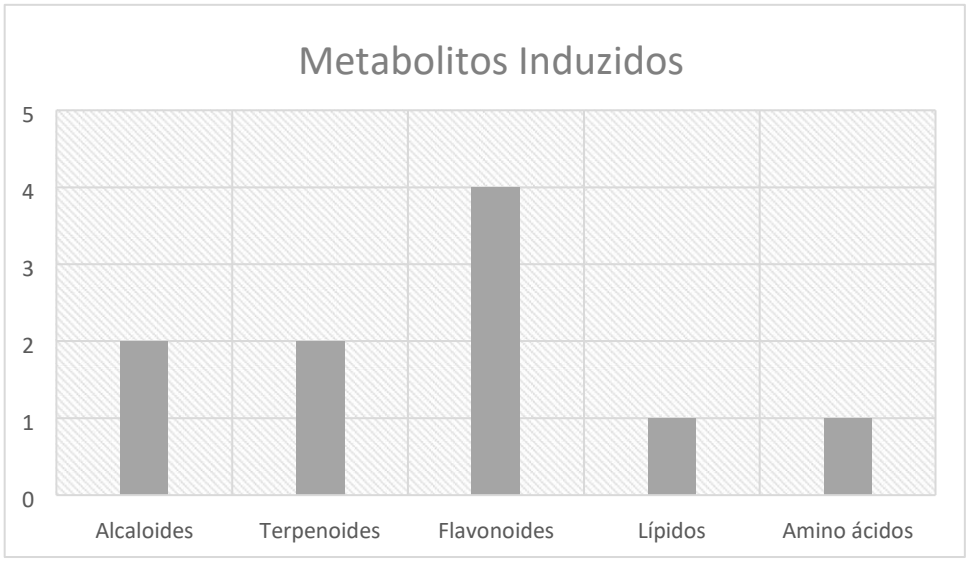

Figura 4.7- Classes de compostos biológicos induzidos identificados na raíz de sobreiro após micorrização com o fungo Pisolithus tinctorius

Os flavonoides foi a classe de compostos mais induzida após a micorrização.

No que respeita aos metabolitos reprimidos na raíz de sobreiro após simbiose com o fungo micorrízico (Fig. 4.8), foram identificadas 6 classes diferentes de compostos, incluindo lípidos, amino ácidos, açúcares e 3 classes de compostos fitoquímicos, fenilpropanoides, flavonoides e alcaloides. A classe com a maior quantidade de metabolitos reprimidos foi a dos lípidos e, dentro dos compostos fitoquímicos, a dos alcaloides.

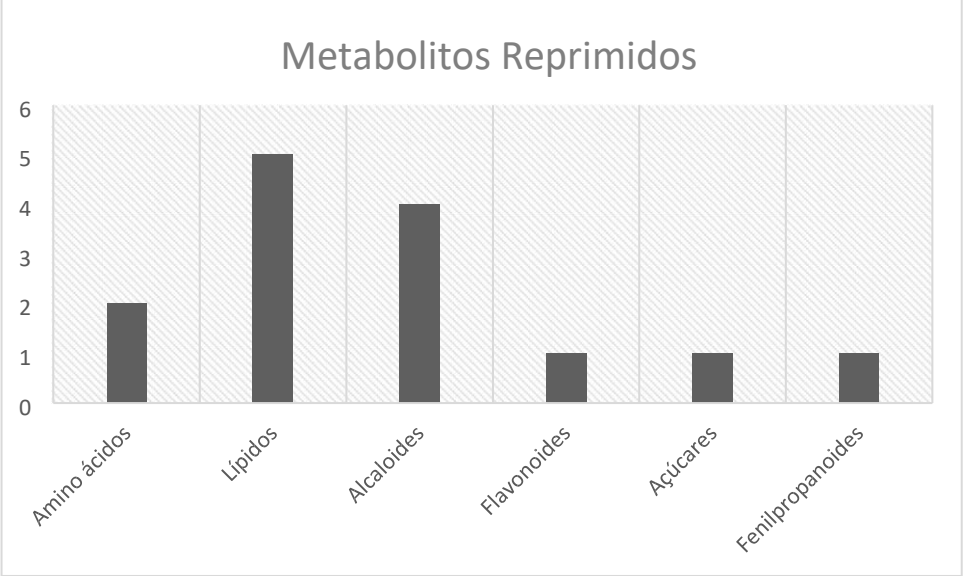

Figura 4.8 - Classes de compostos biológicos reprimidos identificados na raíz de sobreiro após micorrização com o fungo Pisolithus tinctorius

4.2 Método voom-limma

Este método foi utilizado no presente estudo como uma tentativa de possibilitar uma alternativa à identificação de compostos diferencialmente acumulados.

No processo desta análise, foram avaliadas duas estratégias, sendo estas: a utilização do limma com (i) varFilter + voom e (ii) apenas voom. Em qualquer uma das opções, os valores da estatística B tiveram um comportamento idêntico para todos os compostos, pelo que optámos por mostrar como exemplo apenas o composto MeOH com ESI Positivo para a opção (i) (tabela 4.1). Recorrendo à visualização do respetivo Volcano Plot, sendo este habitualmente usado para apresentar resultados de sequências de RNA ou experiências de ómicas. Trata-se de um gráfico de dispersão que mostra a significância estatística *versus* magnitude da acumulação diferencial. Neste gráfico, os genes mais induzidos estão mais para a direita, os reprimidos mais para a esquerda e os mais estatisticamente significativos estão mais para o topo. (Doyle, M., 2019; Batut et al., 2018). Verifica-se também que não tem sequer o aspeto de um vulcão (termo que dá origem à designação do gráfico), pelo facto dos valores de B serem praticamente iguais (Figura 4.9). Apesar dos valores-p ajustados serem inferiores aos níveis de significâncias usuais, não se considera a estatística t-moderada na tomada de decisão por não serem coerentes com os resultados obtidos pela estatística B.

Tabela 4.1 - Resultados relativos ao limma com Voom e varFilter para uma das componentes (MeOH Positivo)

massas	t	P.Value	adj.P.Val	B
362.338335	-92.81102628	1.008349796e-08	7.7340429e-06	-4.59511976055794
317.1196933	-75.80571612	2.562732830e-08	8.1188124e-06	-4.59511976058704
425.2858033	-69.96390826	3.70838950e-08	8.1188124e-06	-4.59511976060224
400.61096	67.97962238	4.234061257e-08	8.1188124e-06	-4.59511976060831

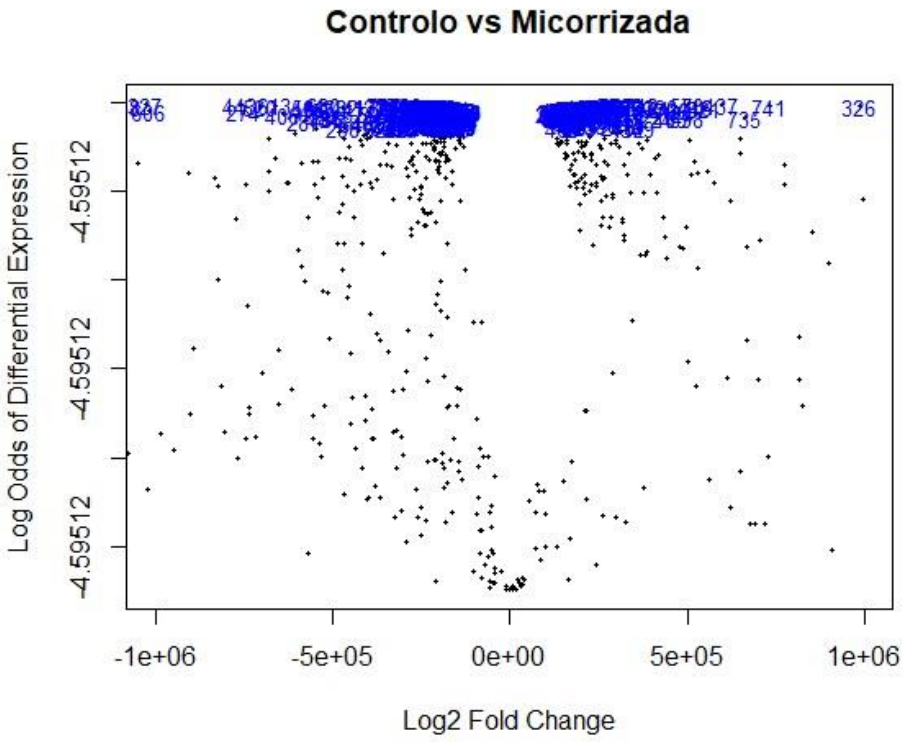

Figura 4.9 – Volcano Plot referente à análise apresentada na Tabela 4.1

Uma das possibilidades dos resultados obtidos não terem sido os melhores, é o facto da função voom ter sido desenvolvida para transformar dados de sequenciação onde surgem muitos zeros, o que acontece com menor frequência nos dados aqui tratados. Por outro lado, o método por trás do *package* limma pressupõe um modelo normal subjacente aos dados, que será um pressuposto difícil de validar com amostras tão pequenas. Os resultados obtidos com a opção (i) varFilter + voom e com a opção (ii) apenas voom, não serão usados para tirar conclusões.

4.3 O PLS-DA e o uso do MetaboAnalyst

O *software* MetaboAnalyst (https://www.metaboanalyst.ca), que permite identificar os metabolitos que se expressam diferencialmente em duas condições distintas (por ex. plants Micorrizadas vs Controlo) e é muito utilizado pelos cientistas que trabalham com dados de metabolómica.

Através deste programa foi possível calcular o VIP (uma medida da importância de uma variável no modelo PLS-DA) e tentar compreender se eventualmente os resultados obtidos se aproximariam dos conseguidos pelo método *Rank Products.*

Através do uso do PLS-DA obtiveram-se estes valores, e visto que a média quadrática dos VIP scores é igual 1, a regra de maior que 1 é geralmente usado como critério para seleção de variáveis. (Chong, I-G. & Jun, C-H.; 2005)

Com o método do PLS-DA, a quantidade de massas obtidas através dos valores VIP foi a seguinte:

- Fase Orgânica com ESI(-) selecionaram-se 5 massas que apresentavam um VIP>1
- Fase Orgânica com ESI(+) selecionaram-se 122 massas que apresentavam um VIP>1

- ACN com ESI(-) selecionaram-se 22 massas que apresentavam um VIP>1
- ACN com ESI(+) selecionaram-se 54 massas que apresentavam um VIP>1

- MeOH com ESI(-) selecionaram-se 44 massas que apresentavam um VIP>1
- MeOH com ESI(+) selecionaram-se 155 massas que apresentavam um VIP>1

- H2O com ESI(-) selecionaram-se 35 massas que apresentavam um VIP>1
- H2O com ESI(+) selecionaram-se 117 massas que apresentavam um VIP>1

Por outro lado, temos os dados com um alinhamento a 1ppm, que permitiu uma comparação mais precisa entre o PLS-DA (VIP>1) e o *Rank Products* (FDR<0.1 e FDR<0.2). Os resultados seguintes mostram o número de compostos discriminantes (massas) selecionadas através do PLS-DA com VIP>1, do *Rank Products* com (FDR<0.1 e FDR<0.2) e as que se encontram em comum entre ambos os métodos e que respeitam a condição do VIP e as condições do *Rank Products*:

Tabela 4.2 - Número de compostos discriminantes (massas) selecionadas através do PLS-DA com VIP>1, Rank Products com (FDR<0.1 e FDR<0.2) e os compostos discriminantes em comum.

Solvente e Métodos de Ionização	PLS-DA (VIP>1)	*Rank Products* (FDR<0.1)	*Rank Products* (FDR<0.2)	Compostos discriminantes em comum
Fase Orgânica com ESI(-)	5	3	5	0
Fase Orgânica com ESI(+)	122	55	95	36
ACN com ESI(-)	22	12	22	2
ACN com ESI(+)	54	21	30	6
MeOH com ESI(-)	44	24	30	5
MeOH com ESI(+)	155	53	92	26
H2O com ESI(-)	35	12	19	3
H2O com ESI(+)	117	36	58	10

Pode-se observar através dos resultados da tabela acima que as combinações de (Solvente x Métodos de Ionização) onde se verifica maior número de compostos discriminantes em comum são Fase Orgânica com ESI(+) e MeOH com ESI(+), com um total de 36 e 26 massas, respetivamente. Como é de constatar, o método PLS-DA foi o que permitiu identificar uma maior quantidade de massas em comparação com o método do *Rank Products*, quer para FDR<0.1 como para FDR<0.2. Com um valor de FDR<0.2 conseguiu-se identificar um elevado número de massas discriminantes, sendo que Fase Orgânica com ESI(+) e MeOH com ESI(+) continuam a ser as combinações que permitiram identificar o maior numero de massas discriminantes.

Capítulo 5

Discussão e Conclusão

Numa abordagem biológica, no que diz respeito à acumulação diferencial (*fold-change*), os resultados obtidos mostram que a micorrização parece afetar de forma positiva e negativa o mesmo número de metabolitos, o que está de acordo com trabalhos anteriores neste sistema biológico que mostram a mesma tendência para genes e proteínas (Sebastiana et al. 2014; 2017).

Verificou-se que dentro do grupo de metabolitos ativados, ou induzidos, a classe mais representativas foi a dos compostos fitoquímicos que inclui metabolitos secundários de origem vegetal, que têm variadas funções incluindo, defesa contra agentes patogénicos e insetos, proteção contra stresses ambientais (ex. seca, excesso de luz, radiação U.V., etc), atração de insetos polinizadores (aromas e pigmentos das flores), desenvolvimento e sinalização entre plantas e microrganismos. A classe de compostos fitoquímicos mais induzida após a micorrização foi a dos flavonoides, o que se encontra de acordo com o papel atribuído a estes compostos na sinalização entre plantas e microrganismos, incluindo microrganismos simbióticos, como é o caso da interação entre o sobreiro e fungos micorrízicos do solo. Vários estudos revelaram que os flavonoides são utilizados pelas plantas na defesa contra patógenos e herbívoros (Hölscher et al. 2016) e nas interações entre as plantas e microrganismos simbióticos, como os nódulos fixadores de azoto e as micorrizas das espécies hortícolas (Hassan & Mathesius, 2012). Nas simbioses das plantas, os flavonoides atuam como quimioatractores, indutores de genes específicos simbióticos e reguladores do desenvolvimento da raíz simbiótica (Hassan & Mathesius, 2012). Refira-se que em trabalhos anteriores sobre as micorrizas de sobreiro, os genes envolvidos no metabolismo dos flavonoides foram dos mais ativados na raíz (Sebastiana et al. 2014). Estes resultados sugerem que nas micorrizas das espécies florestais, como o sobreiro, os flavonoides também estão implicados na comunicação entre planta e fungo simbiótico. Um estudo mais pormenorizado dos flavonoides identificados irá permitir avanços no conhecimento dos processos de reconhecimento entre as plantas florestais e os fungos micorrízicos.

No grupo dos metabolitos reprimidos, os compostos lipídicos identificados pertenciam maioritariamente à classe dos ácidos gordos e incluíam por exemplo o ácido palmítico, o ácido oleico e o ácido laurico. Sabe-se que as plantas produzem uma enorme variedade de ácidos gordos que constituem uma parte importante das membranas celulares e são acumulados como compostos de reserva de energia (ex. sementes oleaginosas). A diminuição dos níveis de ácidos gordos nas raízes micorrizadas de sobreiro detetada no presente trabalho irá ser investigada.

Relativamente aos métodos estatísticos considerados, uma vez que a função voom apenas permite normalizar dados de sequenciação onde surgem vários zeros, facto já anteriormente mencionado e que não se adequa aos dados aqui analisados, apontamos este facto como um dos motivos para os resultados não estarem de acordo com os obtidos pelo procedimento usual (PLS-DA).

No entanto, este mesmo facto, o de não haver zeros, permitiu que o *Rank Products* operasse eficazmente já que não houve empates. No seu projeto de mestrado, Catarina Almeida (2013) verificou que o *package* RankProd não poderia ser aplicado a dados de sequenciação precisamente por haver muitos zeros e o método não conseguir lidar com tantos empates.

Através do MetaboAnalyst, conseguiu-se identificar um número mais elevado de massas diferencialmente expressas recorrendo-se ao PLS-DA. Em contraste, para o *Rank Products* este não foi o caso, as massas diferencialmente expressas obtidas foram bastante inferiores, podendo isto ser indício de que a normalização do voom pode não ter sido ser a mais apropriada tendo em conta os dados em causa.

Em termos de filtragens, tanto para o PLS-DA como para o *Rank Products* (através da função varFilter) é idêntica, mas a normalização é diferente. Como já foi referido anteriormente, a função voom não faz uma normalização adequada a estes dados e por isso devem ser tentadas outras abordagens.

Na comparação entre o *Rank Products* e o PLS-DA, o primeiro acaba por ser menos vantajoso que o segundo, apesar do primeiro ser fácil de aplicar, ser intuitivo, produzir *outputs* em vários formatos como tabelas e gráficos (com pontos de corte diversos), e permitir também o cálculo da razão das falsas descobertas. Com o segundo é que se selecionou mais e diferentes massas, apesar de como já foi referido, o segundo contém uma estrutura interna de funcionamento que nem sempre é clara.

Referências

Adriaensen, K., Lelie D., Laere A., Vangronsveld, J., & Colpaert J.V. (2003). A zinc-adapted fungus protects pines from zinc stress. *New Phytologist*, 161 (2), 549–555. DOI: 10.1046/j.1469-8137.2003.00941.x

Almeida, C. (2013). Study of the Role of SETD2 Mutations in clear cell Renal Cell Carcinoma (ccRCC). Projeto de Mestrado em Bioestatística, Universidade de Lisboa.

Barea, J.M., Palenzuela, J., Cornejo, P., Sánchez-Castro, I., Navarro-Fernández, C., LópezGarcía, A., Estrada, B., Azcón, R., Ferrol, N., & Azcón-Aguilar, C. (2011). Ecological and functional roles of mycorrhizas in semiarid ecosystems of Southeast Spain. *Journal of Arid Environments*, 75 (12), 1292–1301. DOI: 10.1016/j.jaridenv.2011.06.001

Batut et al. (2018). Community-Driven Data Analysis Training for Biology. *Cell System*s, 6(6) ,752-758. DOI : 10.1016/j.cels.2018.05.012

Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 107(21), 9546–9551. DOI:10.1073/pnas.0914005107

Brazanti, M.B., Rocca, E., & Pisi, E. (1999). Effect of ectomycorrhizal fungi on chestnut ink disease. *Mycorrhiza*, 9(2), 103–109. DOI:10.1007/s005720050007

Breitling, R., Armengaud, P., Amtmann, A., & Herzyk, P. (2004). Rank Products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573, 83-92. DOI:10.1016/j.febslet.2004.07.055

Brundrett, M. C. (2002). Coevolution of roots and mycorrhizas of land plants. *New Phytologist*, 154 (2), 275-304. DOI: 10.1046/j.1469-8137.2002.00397.x

Chong, J., Soufan, O., Li, C., Caraus, I., L. S., Bourque, G., Wishart, D.S., & Xia, J. (2018). MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic acids research*, 46 (W1),W486-W494. DOI: 10.1093/nar/gky310

Denkert C., Budczies J., Kind T., Weichert W., Tablack P., Sehouli J., Niesporek S., Konsgen D., Dietel M., & Fiehn O. (2006). Mass spectrometry-based metabolic profiling reveals different metabolite patterns in invasive ovarian carcinomas and ovarian borderline tumors. *Cancer Research*, 66(22), 10795–10804. DOI: 10.1158/0008-5472.CAN-060755

Doyle, M. (2019). Visualization of RNA-Seq results with Volcano Plot (Galaxy Training Materials). Consultado em 28 Fev. 2020. Disponível em https://galaxyproject.github.io/training-material/topics/transcriptomics/tutorials/rna-seqviz-with-volcanoplot/tutorial.html

Duñabeitia M.K., Hormilla S., Garcia-Plazaola J.I., Txarterina K., Arteche U., & Becerril J.M. (2004). Differential responses of three fungal species to environmental factors and their

role in the mycorrhization of Pinus radiata D. Don. *Mycorrhiza,* 14(1), 11-18. DOI: 10.1007/s00572-003-0270-5

Dunkler, D., Sanchez-Cabo, F. & Heinze, G. (2011). Statistical Analysis Principles for Omics Data. *Methods in molecular biology (Clifton, N.J.)*, 719, 113-31. DOI:10.1007/978-161779-027-0_5.

Eisinga, R., Breitling, R., Heskes, T. (2013). The exact probability distribution of the rank product statistics for replicated experiments. *FEBS Lett* ,587(6), 677-82. DOI: 10.1016/j.febslet.2013.01.037

Eriksson, L., Johansson, E., Kettapeh-Wold, S., & Wold, S. (1999). *Introduction to Multi- and Megavariate Data Analysis using Projection Methods (PCA & PLS)*. Umeå: Umetrics.

Fiehn O., Kloska S., & Altmann T. (2001). Integrated studies on plant biology using multiparallel techniques. *Current Opinion in Biotechnology*, 12(1), 82–86. DOI: 10.1016/S09581669(00)00165-8

Fini, A., Frangi, P., Amoroso, G., Piatti, R., Faoro, M., Bellasio, C., & Ferrini, F. (2011). Effect of controlled inoculation with specific mycorrhizal fungi from the urban environment on growth and physiology of containerized shade tree species growing under different water regimes. *Mycorrhiza*, 21(8), 703-719. DOI: 10.1007/s00572-011-0370-6

Flores-Monterroso, A. & Canales, J. & de la torre, F. & Avila, C. & Cánovas, F. (2013). Identification of genes differentially expressed in ectomycorrhizal roots during the Pinus pinaster–Laccaria bicolor interaction. *Planta*, 237, DOI:10.1007/s00425-013-1874-4

Gamache P.H., Meyer D.F., Granger M.C., & Acworth I.N. (2004) Metabolomic applications of electrochemistry/mass spectrometry. *American Society for Mass Spectrometry*, 15(12), 1717–1726. DOI: 10.1016/j.jasms.2004.08.016

Gentleman, R., Carey, V., Huber, W., Hahne, F. (2019). genefilter: genefilter: methods for filtering genes from high-throughput experiments. R package version 1.66.0.

Goodacre R., Vaidyanathan S., Dunn W.B., Harrigan G.G., & Kell D.B. (2004). Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends in Biotechnology*, 22(5), 245–252. DOI: 10.1016/j.tibtech.2004.03.007

Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., & Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*, 879, 10-23, DOI:10.1016/j.aca.2015.02.012

Han, J., Danell, R. M., Patel, J. R., Gumerov, D. R., Scarlett, C. O., Speir, J. P., & Borchers, C. H. (2008). Towards high-throughput metabolomics using ultrahigh-field Fourier transform ion cyclotron resonance mass spectrometry. Metabolomics. *Official journal of the Metabolomic Society*, 4(2), 128–140. DOI:10.1007/s11306-008-0104-8

Hassan, H. & Mathesius, U. (2012). The role of flavonoids in root–rhizosphere signalling: opportunities and challenges for improving plant–microbe interactions. *Journal of Experimental Botany*, 63(9), 3429–3444. DOI: 10.1093/jxb/err430

Hölscher, D., Buerkert, A., Schneider, B. (2016). Phenylphenalenones accumulate in plant tissues of two banana cultivars in response to herbivory by the banana weevil and banana stem weevil. *Plants*, 5, 34. DOI: 10.3390/plants5030034.

Hong, F., Breitling, R., McEntee, C.W., Wittner, B.S., Nemhauser, J.L., & Chory, J. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in metaanalysis. *Bioinformatics*, 22(22), 2825-2827. DOI:10.1093/bioinformatics/btl476

Il-Gyo, C., Chi-Hyuck, J. (2005). Performance of some variable selection methods when multicollinearity is present. Chemometr Intell Lab Syst 78: 103-112. *Chemometrics and Intelligent Laboratory Systems*, 78, 103-112. DOI: 10.1016/j.chemolab.2004.12.011.

Johansson, E., Garg, P. & Burgers, P.M.G. (2004). The Pol32 Subunit of DNA Polymerase d Contains Separable Domains for Processive Replication and Proliferating Cell Nuclear Antigen (PCNA) Binding. *The Journal of Biological Chemistry*, 279, 1907-1915. DOI: 10.1074/jbc.M310362200

Koziol, J.A. (2010). Comments on the rank product method for analyzing replicated experiments, *FEBS Letters*, 584(5), 941-944. DOI:10.1016/j.febslet.2010.01.031

Krishnan, A., Williams, L.J., McIntosh, A.R., & Abdi, H. (2011). Partial least squares (PLS) methods for neuroimaging: a tutorial and review. *Neuroimage* 56(2), 455-75. DOI: 10.1016/j.neuroimage.2010.07.034

Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), R29. DOI: 10.1186/gb-2014-15-2-r29

Lönnstedt, I. & Speed, T. (2001). Replicated Microarray Data. Statistica Sinica, 12(1)

Maia, M., Monteiro, F., Sebastiana, M., Marques, A.P., António E.N. Ferreira, Freire, A.P., Cordeiro, C., Figueiredo, A., & Silva, M.S. (2016). Metabolite extraction for highthroughput FTICR-MS-based metabolomics of grapevine leaves, *EuPA Open Proteomics*, 12, 4-9, DOI:10.1016/j.euprot.2016.03.002

Mehmood, T., Liland, K.H., Snipen, L., & Saebo, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118, 62–69. DOI: 10.1016/j.chemolab.2012.07.010

Mehmood, T., Martens, H., Sæbø, S., Warringer, J., & Snipen, L. (2011). A partial least squares based algorithm for parsimonious variable selection. *Algorithms Molecular Biology 6*, 27. DOI:10.1186/1748-7188-6-27

Misra, B., Langefeld, C., Olivier, M., & Cox, L. (2019). Integrated Omics: Tools, Advances, and Future Approaches. *Journal of Molecular Endocrinology*, 62(1), R21–R45. DOI:10.1530/JME-18-0055.

Nehls, U., Grunze, N., Willmann, M., Reich, M., & Kuster, H. (2007). Sugar for my honey: carbohydrate partitioning in ectomycorrhizal symbiosis. *Phytochem*, 68(1), 82–91. DOI: 10.1016/j.phytochem.2006.09.024

Nicholson, J. K., Lindon, J. C. & Holmes, E. (1999). 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29(11), 11811189, DOI: 10.1080/004982599238047

Núñez, J.A.D., Serrano, J.S., Barreal J.A.R., & Gonzáles, J.A.S.O. (2006). The influence of mycorrhization with Tuber melanosporum in the afforestation of a Mediterranean site with Quercus ilex and Quercus faginea. *Forest Ecology and Management*, 231(1), 226233. DOI: 10.1016/j.foreco.2006.05.052. 3

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1), 45–50. doi:10.4103/0301-4738.37595

Pérez-Enciso, M., & Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human Genetics*, 112(5-6), 581–592. DOI: 10.1007/s00439-003-0921-9

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Reiner, A., Yekutieli, D. & Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3), 368–375. DOI:doi.org/10.1093/bioinformatics/btf877

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research,* 43 (7), e47. DOI:10.1093/nar/gkv007

Sebastiana, M., Martins, J., Figueiredo, A., Monteiro, F., Sardans, J., Penuelas, Josep, Silva, A., Roepstorff, P., Pais, M., & Coelho, A. (2016). Oak protein profile alterations upon root colonization by an ectomycorrhizal fungus. *Mycorrhiza*, 27(2), 110. DOI: 10.1007/s00572-016-0734-z

Sebastiana, M., Pereira, V.T., Alcântara, A., Pais, M.S., & Silva, A.B. (2013). Ectomycorrhizalinoculation with Pisolithus tinctorius increases the performance of Quercus suber L. (cork oak) nursery and field seedlings. *New forests*, 44, 937-949. DOI: 10.1007/s11056-013-9386-4

Sebastiana, M., Vieira, B., Lino-Neto, T., Monteiro, F., Figueiredo, A., Sousa, L., Pais, M.S., Tavares, R., & Paulo, O.S. (2014). Oak root response to ectomycorrizhal symbiosis establishment: RNA-Seq derived transcript identification and expression profiling. *PloS ONE*, 9(5), e98376. DOI: 10.1371/journal.pone.0098376

Smith, S.E. & Read, D.J. (1997) *Mycorrhizal Symbiosis*. Academic Press, London.

Smyth, G. & Altman, N. (2013). Separate-channel analysis of two-channel microarrays: Recovering inter-spot information. *BMC bioinformatics*, 14, 165. DOI:10.1186/14712105-14-165.

Smyth G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, *3*, Article3. DOI: 10.2202/1544-6115.1027

Southworth, D., Carrington, E.M., Frank, J.L., Gould, P., Harrington C.A., & Devine W.D. (2009). Mycorrhizas on nursery and field seedlings of Quercus garryana. *Mycorrhiza*, 19(3), 149–158. DOI: 10.1007/s00572-008-0222-1

Tenenhaus, M. (1998). *La Régression PLS*. Paris: Editions Technip.

Wang B., & Qiu, Y.L. (2006) Phylogenetic distribution and evolution of mycorrhizas in land plants. *Mycorrhiza*, 16(1), 299–363. DOI: 10.1007/s00572-005-0033-6

Want E., & Masson P. (2011). Processing and Analysis of GC/LC-MS-Based Metabolomics Data. *Metabolic Profiling. Methods in Molecular Biology (Methods and Protocols)*, 708, 277-298. DOI: 10.1007/978-1-61737-985-7_17

Xia, J., Psychogios, N., Young, N., & Wishart, D. S. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic acids research*, 37(W), W652–W660. DOI:10.1093/nar/gkp356

Apêndice

Outputs do Rank Products

Tabelas de ACN

Tabela A1 - Massas diferencialmente acumuladas para ACN negativo (Down)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
4	1.817	0.105	0.0007957	0.000004573
146	2.621	0.1626	0.002239	0.00002574
5	4.762	0.1846	0.01899	0.0003274
96	8.291	0.2309	0.1084	0.002491
38	9.967	0.2444	0.1585	0.004553
37	10.84	0.3012	0.1725	0.005948
141	11.41	0.2593	0.1735	0.006978
131	13.3	0.2794	0.2424	0.01115
118	13.33	0.2727	0.217	0.01122
94	13.39	0.2796	0.1979	0.01137

Tabela A2 – Massas diferencialmente acumuladas para ACN positivo (Down)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
106	2.714	0.1159	0.001764	0.000005879
279	2.884	0.1236	0.001155	0.000007701
71	4.82	0.1499	0.00671	0.0000671
228	5.241	0.04643	0.007053	0.0940
126	5.809	0.1632	0.008496	0.0001416
296	6.952	0.1666	0.01418	0.0002837
113	7.56	0.1549	0.01643	0.0003833
251	8.005	0.1768	0.0176	0.0004694
205	10.26	0.1851	0.03659	0.001098
202	10.66	0.1628	0.03741	0.001247
269	12.56	0.2215	0.05791	0.002123
271	13.63	0.2247	0.06881	0.002753
26	15.55	0.2387	0.09571	0.004147
231	18.81	0.249	0.1577	0.007361
285	19.28	0.2644	0.1584	0.007919

Tabela A3 – Massas diferencialmente acumuladas para ACN negativo (Up)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
145	1	18.7	0.00003303	0.0000001898
140	2.52	9.612	0.001874	0.00002154
167	3.175	7.925	0.003481	0.00006002
143	4.217	6.682	0.008662	0.0001991
15	4.762	6.853	0.01139	0.0003274
110	6	5.018	0.0234	0.000807
109	7.612	4.515	0.04636	0.001865
166	8.811	4.804	0.06631	0.003049
148	9.655	4.052	0.07947	0.004111
171	11.89	3.729	0.1379	0.007925
165	13.51	3.479	0.1847	0.01168
119	13.98	3.48	0.1875	0.01293

Tabela A4 – Massas diferencialmente acumuladas para ACN positivo (Up)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
125	2.621	11.38	0.001506	0. 000005021
252	4.642	10.31	0.008634	0.000005576
145	5.474	10.25	0.01118	0.0001118
186	9.205	9.642	0.0571	0.0007613
194	10.3	6.268	0.06667	0.001111
291	11.66	5.801	0.08353	0.001671
295	12.78	5.522	0.0963	0.002247
138	13.52	5.424	0.1006	0.002684
117	13.99	6.851	0.09959	0.002988
88	14.79	5.331	0.1066	0.003555
240	14.95	5.127	0.1004	0.00368
112	17.62	4.643	0.1517	0.006066
221	19.12	4.583	0.1785	0.007733
246	19.98	4.532	0.1886	0.0088
109	20.08	4.935	0.1786	0.008928

Tabelas de H2O

Tabela A5 – Massas diferencialmente acumuladas para H2O negativo (Down)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
112	1.587	0.0529	0.0007556	0.000005949
110	2	0.03261	0.001193	0.00001879
103	2.621	0.06089	0.002802	0.00006619
124	4.38	0.09984	0.01899	0.0005983
36	5.944	0.1158	0.0489	0.001925
122	6.868	0.1252	0.06741	0.003185
93	8	0.1337	0.09629	0.005307
113	8.896	0.1421	0.1188	0.007485
81	8.996	0.1344	0.1094	0.007756
114	9.39	0.1392	0.1128	0.008886

Tabela A6 – Massas diferencialmente acumuladas para H2O positivo (Down)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
451	3.175	0.05418	0.0006584	0.0000009501
685	3.302	0.06655	0.00039	0.000001125
422	5.313	0.07349	0.001861	0.000008056
519	5.485	0.04901	0.001585	0.000009146
608	6.214	0.08361	0.002078	0.000001499
394	7.731	0.09518	0.004044	0.000003501
680	8.759	0.1041	0.005588	0.000005644
639	9.166	0.1044	0.005768	0.000006658
254	11.08	0.111	0.009996	0.0001298
42	11.17	0.07714	0.009264	0.0001337
212	16.54	0.1324	0.03118	0.0004949
473	20.71	0.1129	0.05836	0.001011
575	20.78	0.148	0.05444	0.001021
381	22.23	0.1529	0.06231	0.001259
166	22.88	0.1429	0.06353	0.001375
68	24.97	0.1276	0.07783	0.001797
536	26.42	0.1632	0.08686	0.002131
657	27.85	0.1676	0.09601	0.002494
362	27.95	0.1378	0.09196	0.002521
521	28.31	0.1652	0.09069	0.002617
625	28.74	0.1632	0.09036	0.002738
377	29.41	0.1682	0.09235	0.002932
635	29.76	0.1442	0.09151	0.003037
370	31.34	0.1874	0.102	0.003534
453	31.8	0.1665	0.1022	0.003687
661	32.93	0.1869	0.1087	0.00408
30	33.05	0.1879	0.1058	0.004124
278	34.27	0.1899	0.1134	0.004581
392	35.58	0.1993	0.1219	0.005101
137	36.03	0.1875	0.1221	0.005285
445	37.72	0.1759	0.1346	0.006021
537	40.74	0.1739	0.1619	0.007477
592	41.8	0.2054	0.1687	0.008034

175	42.21	0.1917	0.1682	0.008254
630	42.5	0.2104	0.1666	0.008413
341	43.54	0.2155	0.1732	0.008995
497	44.22	0.2184	0.1758	0.009386
368	46.11	0.2231	0.1919	0.01052
491	46.72	0.2129	0.1938	0.0109

Tabela A7 – Massas diferencialmente acumuladas para H2O negativo (Up)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
77	1.26	11.97	0.0002238	0.000001762
79	2	11.29	0.001193	0.0001879
78	3	5.334	0.005107	0.0001206
56	5.944	3.802	0.06113	0.001925
37	6.162	4.818	0.05555	0.002187
115	9.166	3.149	0.1742	0.008231
111	9.221	2.86	0.1522	0.008389
123	9.59	2.933	0.1507	0.009494
24	10.72	2.646	0.189	0.0134

Tabela A8 – Massas diferencialmente acumuladas para H2O positivo (Up)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
686	5.646	29.81	0.007113	0.00001026
645	6.358	172.9	0.005681	0.0000164
268	6.709	100.2	0.004673	0.00002023
421	9.491	29.74	0.01307	0.00007542
304	12.93	16.1	0.03046	0.0002197
631	18.41	12.87	0.08058	0.0006976
15	18.45	23.9	0.06959	0.000703
234	18.87	18.06	0.06539	0.0007548
4	18.92	21.05	0.05856	0.0007605
662	20.41	11.99	0.06696	0.0009662
669	21.94	11.58	0.07617	0.001209
656	22.77	12.29	0.07831	0.001356
452	23.05	15.4	0.075	0.001407

173	26.45	18.35	0.1057	0.002136
441	28.16	11.24	0.119	0.002576
309	29.74	8.404	0.1313	0.00303
681	30.05	8.322	0.1273	0.003122
612	31.77	9.282	0.1416	0.003678
506	33.87	8.499	0.1615	0.004427

Tabelas de MeOH:

Tabela A9 – Massas diferencialmente acumuladas para MeOH negativo (Down)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
78	1	0.01309	0.00002741	0.0000001435
151	3.107	0.1019	0.003948	0.00004134
145	3.557	0.1229	0.004702	0.00007385
117	5.04	0.07137	0.01486	0.0003112
175	5.313	0.17	0.0147	0.0003848
184	5.518	0.1668	0.01424	0.0004474
87	7.306	0.1848	0.03392	0.001243
167	8	0.2245	0.04056	0.001699
179	8.378	0.2199	0.04214	0.001986
174	9.967	0.2503	0.06728	0.003522
187	10.63	0.2552	0.07515	0.004328
24	12.51	0.2797	0.115	0.007223
160	12.83	0.2809	0.1145	0.007796
176	14.23	0.2947	0.1454	0.01066
185	16.01	0.318	0.1921	0.01509
168	16.45	0.3035	0.195	0.01633

Tabela A10 – Massas diferencialmente acumuladas para MeOH positivo (Down)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
178	1.71	0.02743	0.00003024	0.00000003942
585	3.557	0.03362	0.0004374	0.00000114
207	3.634	0.04501	0.0003194	0.000001249
146	4	0.04651	0.0003581	0.000001867
240	4.932	0.04925	0.000676	0.000004407
659	6.136	0.05641	0.001344	0.00001052
331	7.612	0.05451	0.002666	0.00002433
100	8.759	0.07079	0.003991	0.00004163
283	9.435	0.06879	0.004676	0.00005487
258	11.55	0.08937	0.008565	0.0001117
429	12.05	0.07745	0.009004	0.0001291
327	12.93	0.09457	0.01049	0.0001642
669	13.33	0.09253	0.01073	0.0001819
606	13.66	0.09548	0.01083	0.0001977
363	14.46	0.08043	0.01223	0.0002392
219	17.61	0.1068	0.02183	0.0004553
575	20.98	0.1168	0.03585	0.0007945
182	22.16	0.1178	0.04012	0.0009415
443	22.57	0.1278	0.04026	0.0009972
274	23.49	0.1297	0.04325	0.001128
652	25.51	0.1253	0.05305	0.001452
290	26.6	0.1303	0.05746	0.001648
520	27.07	0.1383	0.05797	0.001738
740	29.53	0.1427	0.0721	0.002256
613	30.05	0.1469	0.0729	0.002376
716	31.08	0.143	0.07741	0.002624
120	32.12	0.1425	0.08218	0.002893
406	32.43	0.1518	0.08149	0.002975
697	33.2	0.1471	0.08428	0.003187
705	35.55	0.1544	0.09937	0.003887
658	36.56	0.1601	0.1042	0.004212
641	37.18	0.1555	0.106	0.004423

287	39.33	0.1683	0.1207	0.005193
436	41.27	0.1722	0.1342	0.00595
561	41.57	0.1719	0.133	0.006071
497	44.4	0.1798	0.1555	0.007299
445	45.06	0.174	0.1577	0.007608
661	45.17	0.1775	0.1546	0.007657
677	45.27	0.1763	0.1515	0.007705
693	46.55	0.1785	0.1595	0.008319
630	47.3	0.1855	0.1627	0.008695
484	48.35	0.1853	0.1686	0.009232
90	49.38	0.1792	0.1745	0.009781
736	49.62	0.1866	0.1728	0.009911
466	49.76	0.1883	0.1702	0.009984
587	50.5	0.1671	0.1733	0.01039
208	52.12	0.1843	0.1847	0.01132

Tabela A11- Massas diferencialmente acumuladas para MeOH negativo (Up)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
122	1.442	24.23	0.000203	0.000001063
119	2.289	17.62	0.001001	0.00001048
121	3.53	12.84	0.004555	0.00007154
13	4.642	8.874	0.01065	0.000223
104	5.429	7.757	0.01602	0.0004193
123	5.518	8.606	0.01424	0.0004474
135	6.415	7.578	0.02142	0.0007851
180	8.819	5.447	0.05631	0.002358
190	8.819	5.26	0.05005	0.002358
112	9.524	5.086	0.05805	0.003039
146	10.72	4.735	0.07728	0.004451
171	10.88	4.848	0.07419	0.004661
161	11.97	4.652	0.09254	0.006299
143	13.44	4.335	0.1226	0.008988

Tabela A12 – Massas diferencialmente acumuladas para MeOH positivo (Up)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
383	4.38	25.39	0.002083	0.000002716
432	5.241	30.48	0.002158	0.000005627
720	7.747	16.55	0.006658	0.00002604
635	8.636	16.23	0.007563	0.00003944
442	9.102	21.44	0.007388	0.00004816
660	10.16	13.69	0.009139	0.00007149
645	10.51	15.02	0.008806	0.00008036
426	10.7	15.6	0.008212	0.00008565
391	11.44	14.68	0.009218	0.0001082
485	13.53	11.45	0.01468	0.0001913
527	15.08	13.76	0.01918	0.000275
382	15.29	11.3	0.01838	0.0002876
326	16.58	10.16	0.02212	0.000375
454	17.47	10.3	0.0243	0.0004435
549	20.01	11.37	0.03497	0.0006838
330	23.14	8.814	0.05162	0.001077
741	23.63	7.585	0.05185	0.001149
519	25.27	7.977	0.06015	0.001412
273	25.84	7.999	0.06093	0.001509
735	28.95	6.855	0.0815	0.002125
284	30.26	6.706	0.08861	0.002426
562	30.72	6.725	0.08844	0.002537
689	31.04	7.294	0.08723	0.002616
437	33.18	6.181	0.1017	0.003181
338	35.94	6.879	0.123	0.00401
218	36.35	6.425	0.1223	0.004146
370	37.42	8.438	0.128	0.004506
521	38.84	5.683	0.1372	0.00501
189	39.23	5.965	0.1363	0.005155
576	39.62	5.776	0.1356	0.005303
453	41.5	8.518	0.1495	0.006042
446	41.99	5.428	0.1497	0.006247

161	42.3	5.473	0.1482	0.006377
199	43.08	5.379	0.1514	0.006711
570	44.4	5.182	0.16	0.007302
598	44.75	5.173	0.1589	0.00746
605	46.1	5.228	0.1679	0.008099
55	47.38	6.896	0.1763	0.008737
496	47.66	5.298	0.1746	0.008879
183	47.8	4.841	0.1716	0.008947
721	48.13	5.468	0.1706	0.009121
124	48.84	4.903	0.1733	0.009492
13	49.66	7.229	0.1771	0.009931
706	50.31	4.743	0.1793	0.01029
405	51.78	4.686	0.1896	0.01112

Tabelas da fase Orgânica

Tabela A13 – Massas diferencialmente acumuladas para Org negativo (Down)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
18	1.26	0.3037	0.003756	0.0001212
14	2.884	0.4111	0.1082	0.006979

Tabela A14 – Massas diferencialmente acumualdas para Org positivo (Down)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
373	4.61	0.08412	0.003167	0.000004583
673	5	0.08715	0.0022	0.000006367
346	6.463	0.09062	0.004063	0.00001764
534	6.868	0.09295	0.003862	0.00002236
383	6.952	0.04772	0.003239	0.00002344
276	7.83	0.1025	0.004272	0.00003709
593	8.082	0.1023	0.004136	0.0000419
147	8.387	0.09159	0.004169	0.00004826
505	10.32	0.113	0.007857	0.0001023
399	10.67	0.09548	0.007947	0.000115
680	14.66	0.128	0.02116	0.0003368
164	14.86	0.1263	0.02027	0.0003521
669	15.46	0.1359	0.02132	0.0004011
490	15.72	0.129	0.02092	0.0004238
362	18.47	0.156	0.03274	0.0007108
631	19.15	0.1245	0.03441	0.0007968
356	20.85	0.1413	0.0423	0.001041
312	20.95	0.1576	0.04055	0.001056
539	22.12	0.1629	0.04547	0.00125
637	22.24	0.1566	0.04391	0.001271
542	22.51	0.173	0.04339	0.001319
394	24.1	0.1722	0.05108	0.001626
653	24.53	0.173	0.05154	0.001716
595	26.58	0.1817	0.06292	0.002185
336	27.26	0.1853	0.06515	0.002357
529	28.09	0.1912	0.06856	0.00258
310	29.69	0.185	0.07774	0.003037
517	30.67	0.2025	0.08252	0.003344
240	31.43	0.1999	0.08559	0.003592
604	32.96	0.207	0.09497	0.004123
666	33.71	0.2043	0.09809	0.0044
233	35.05	0.2043	0.1063	0.004924

245	37.18	0.2187	0.122	0.005827
450	39.11	0.2334	0.1366	0.006723
519	39.56	0.2053	0.137	0.00694
451	39.62	0.2301	0.1338	0.006969
582	40.89	0.227	0.1422	0.007614
354	42.72	0.2175	0.1563	0.008596
556	44.09	0.2452	0.1661	0.009376
686	44.6	0.2435	0.1672	0.00968
568	45.77	0.2358	0.1751	0.01039
335	46.03	0.2506	0.1735	0.01055
190	46.22	0.2459	0.1715	0.01067
533	46.5	0.2517	0.1703	0.01085
465	47.65	0.254	0.178	0.01159
564	48.4	0.2457	0.1816	0.01209

Tabela A15 – Massas diferencialmente acumuladas para Org negativo (Up)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
15	1	7.347	0.001041	0.00003357
23	2	3.217	0.02003	0.001292
29	3.302	2.491	0.1264	0.01223

Tabela A16 – Massas diferencialmente acumuladas para Org positivo (Up)

massa.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
654	3.107	50.32	0.0006032	0.0000008729
620	4.273	36.05	0.001159	0.000003355
386	4.82	34.5	0.001265	0.000005491
333	7.23	26.34	0.004714	0.00002729
431	8.052	28.21	0.005706	0.00004129
316	9.182	31.72	0.007781	0.00006757
655	13.22	18.36	0.02358	0.0002389
343	14.08	17.76	0.02545	0.0002946
467	15.39	18.54	0.03036	0.0003954
594	17.91	13.56	0.04455	0.0006447
551	19.05	15.88	0.04924	0.0007839
277	19.26	12.71	0.04671	0.0008111
468	19.63	15.91	0.04581	0.0008618
674	19.81	12.79	0.04378	0.000887
379	20.08	14.03	0.04264	0.0009257
252	20.82	16.21	0.04474	0.001036
679	21.23	11.8	0.04474	0.001101
559	21.46	12.55	0.04368	0.001138
148	22.07	11.19	0.04515	0.001242
374	22.9	11.14	0.04804	0.00139
47	23.15	12.1	0.04732	0.001438
332	23.87	11.58	0.04962	0.00158
347	28.99	9.465	0.08503	0.00283
619	30.78	13.46	0.09722	0.003377
535	31.74	8.254	0.1021	0.003694
230	32.23	8.034	0.1027	0.003866
449	32.86	8.187	0.1046	0.004087
387	35.01	11.79	0.1211	0.004909
392	36.25	11.01	0.1291	0.00542
256	36.53	9.159	0.1276	0.005539
434	37.09	8.139	0.129	0.005787

507	37.42	7.147	0.1282	0.005935
536	37.55	10.47	0.1254	0.00599
167	37.65	11.61	0.1227	0.006036
363	38.31	7.761	0.1252	0.006342
575	39.25	10.16	0.1303	0.00679
48	39.73	9.572	0.1312	0.007025
574	39.81	10.72	0.1285	0.007066
340	42.34	9.927	0.1486	0.008386
334	43.18	9.807	0.153	0.008857
622	44.22	8.858	0.1594	0.009457
638	44.73	6.176	0.1605	0.009754
63	48.17	9.204	0.1918	0.01193
541	48.64	5.828	0.1923	0.01225
220	49.46	6.392	0.1967	0.01281
549	49.49	9.966	0.1928	0.01283
286	49.97	6.831	0.1935	0.01316
538	49.97	9.019	0.1896	0.01317
311	50.86	5.443	0.1946	0.0138

Restantes gráficos de identificação de massas reprimidas e induzidas com pfp<0.2 para H2O, MeOH e Fase Orgânica:

- H2O Positivo com ESI(+):

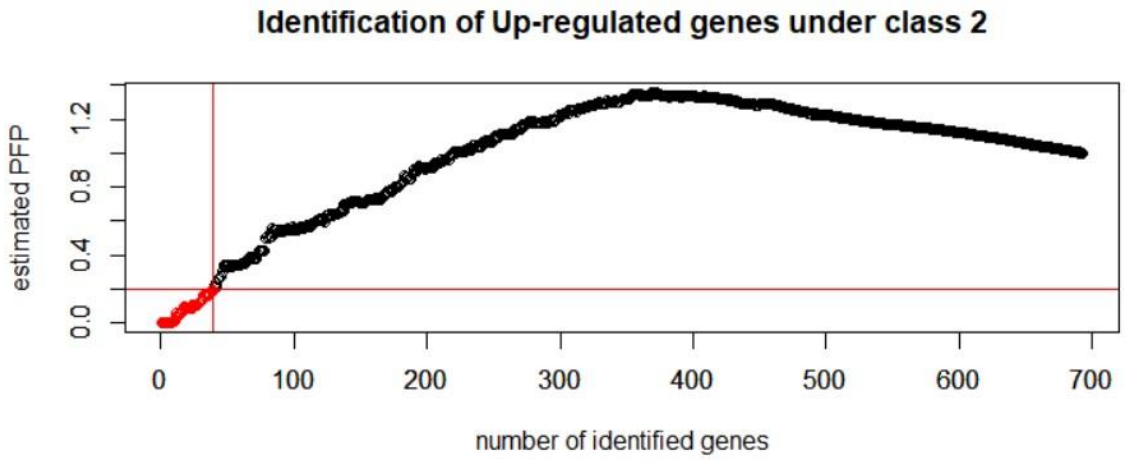

Figura A1 - Gráfico de identificação de massas induzidas com FDR<0.2 para H2O Positivo

- H2O Positivo com ESI(-):

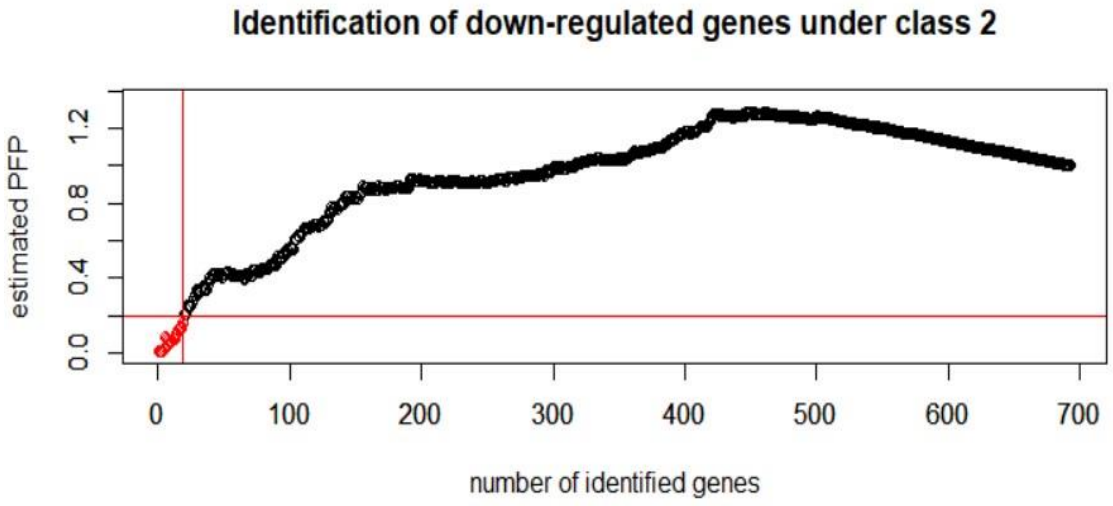

Figura A2 - Gráfico de identificação de massas reprimidas com FDR<0.2 para H2O Positivo

H2O Negativo com ESI(+):

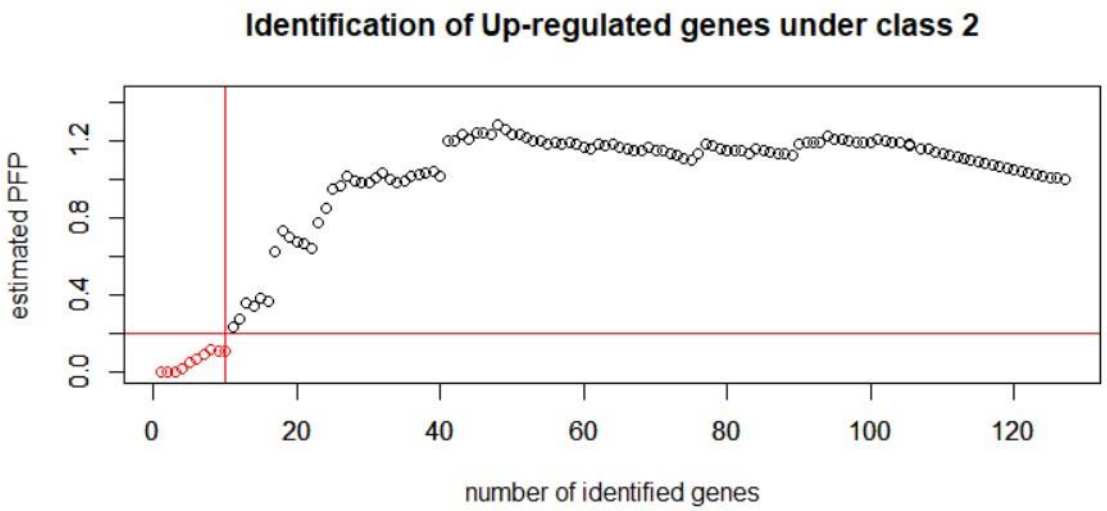

Figura A3 - Gráfico de identificação de massas induzidas com FDR<0.2 para H2O Negativo

- H2O Negativo com ESI(-):

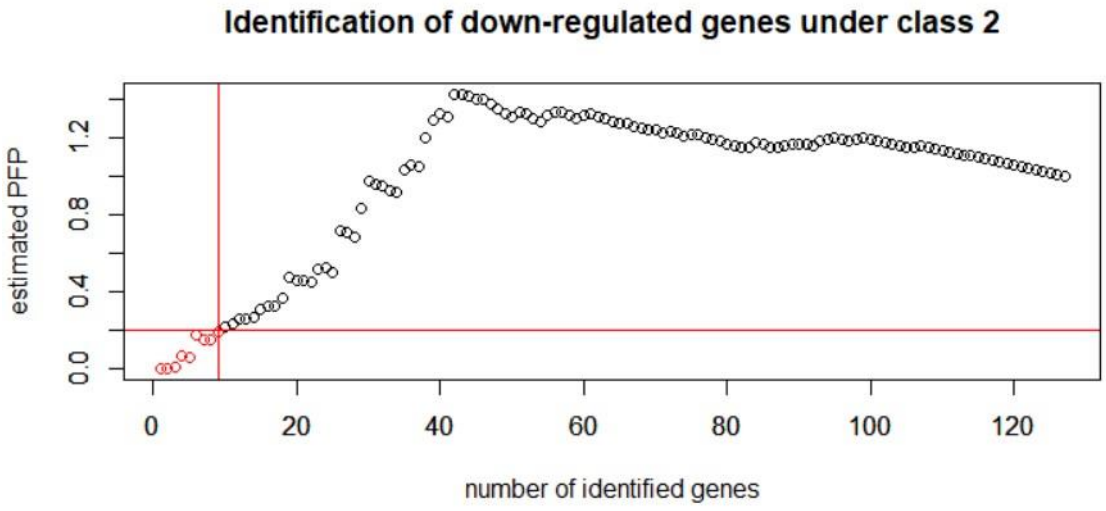

Figura A4 - Gráfico de identificação de massas reprimidas com FDR<0.2 para H2O Negativo

MeOH Positivo com ESI(+):

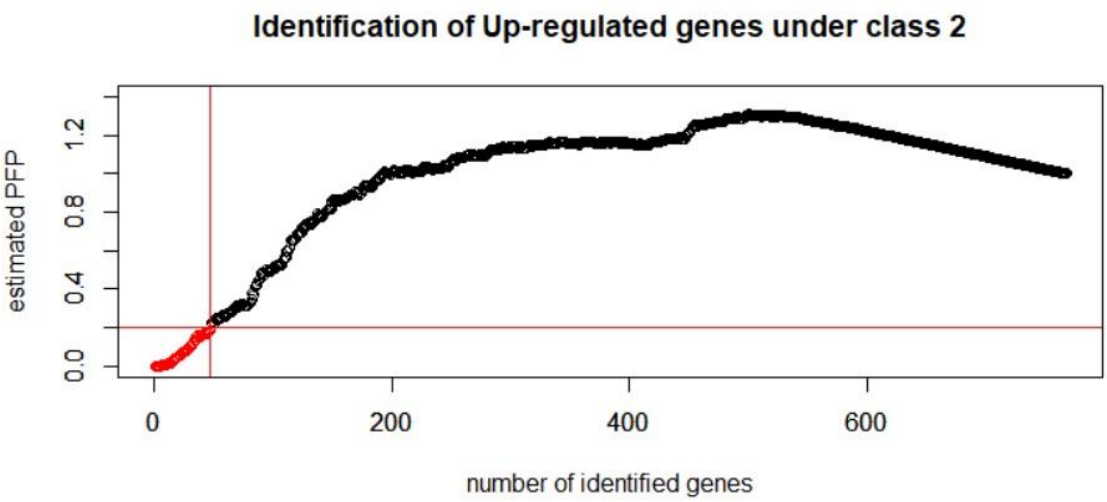

Figura A5 - Gráfico de identificação de massas induzidas com FDR<0.2 para MeOH Positivo

- MeOH Positivo com ESI(-):

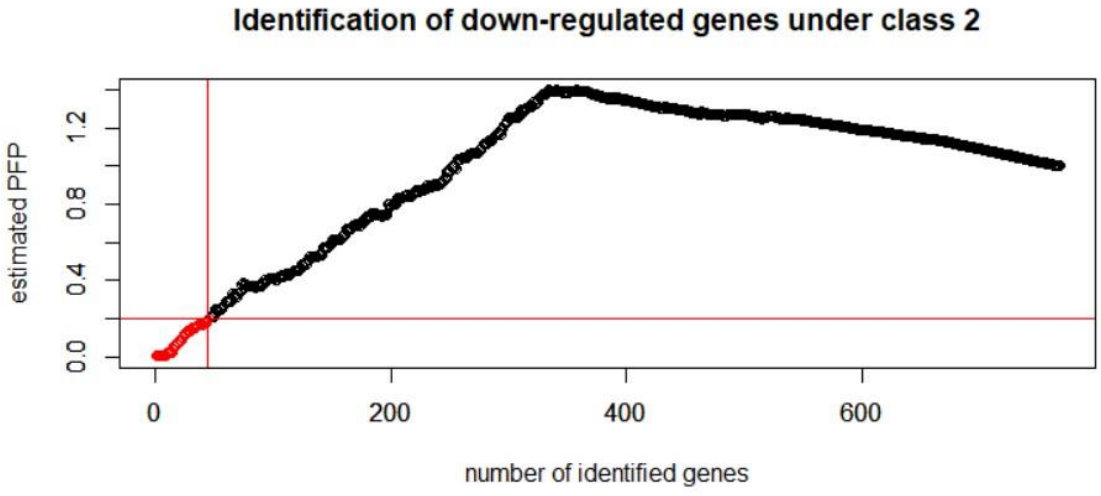

Figura A6 - Gráfico de identificação de massas reprimidas com FDR<0.2 para MeOH Positivo

MeOH Negativo com ESI(+):

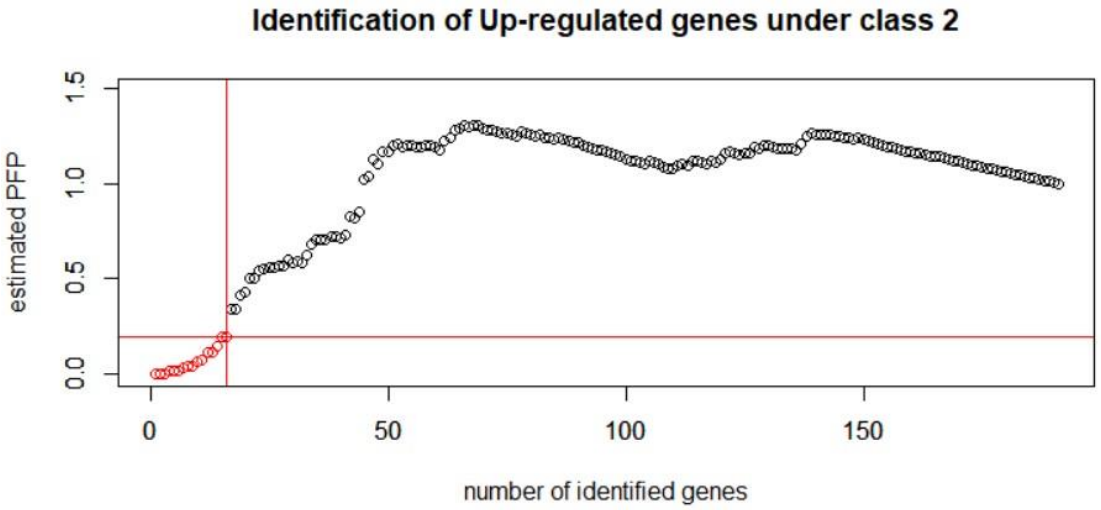

Figura A7 - Gráfico de identificação de massas induzidas com FDR<0.2 para MeOH Negativo

- MeOH Negativo com ESI(-):

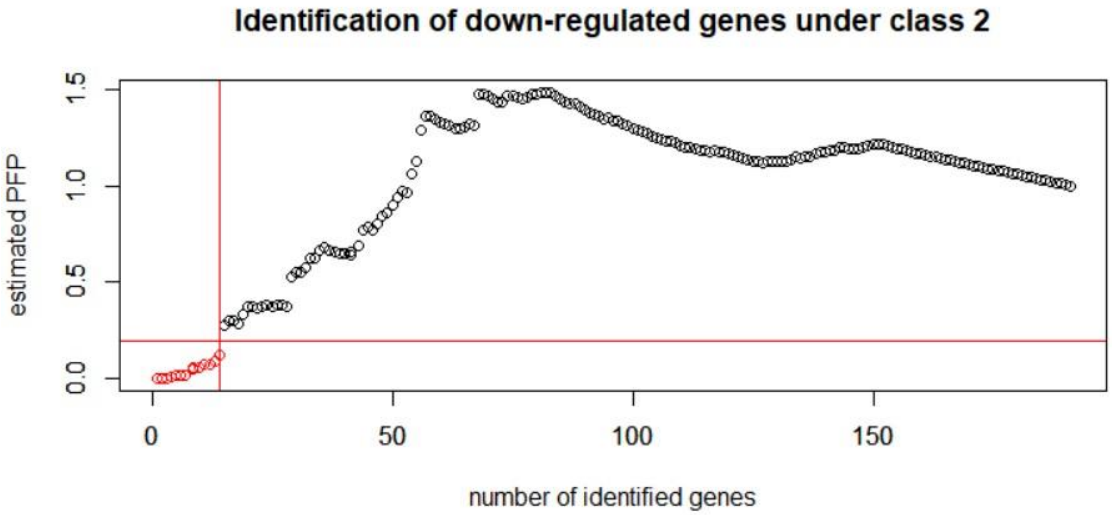

Figura A8 - Gráfico de identificação de massas reprimidas com FDR<0.2 para MeOH Negativo

Fase Orgânica Positivo com ESI(+):

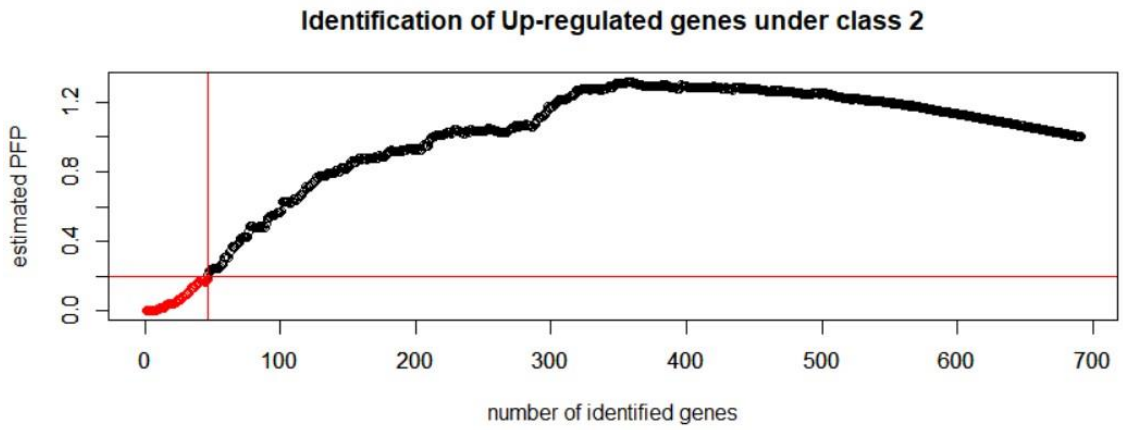

Figura A9 - Gráfico de identificação de massas induzidas com FDR<0.2 para Fase Orgânica Positivo

- Fase Orgânica Positivo com ESI(-):

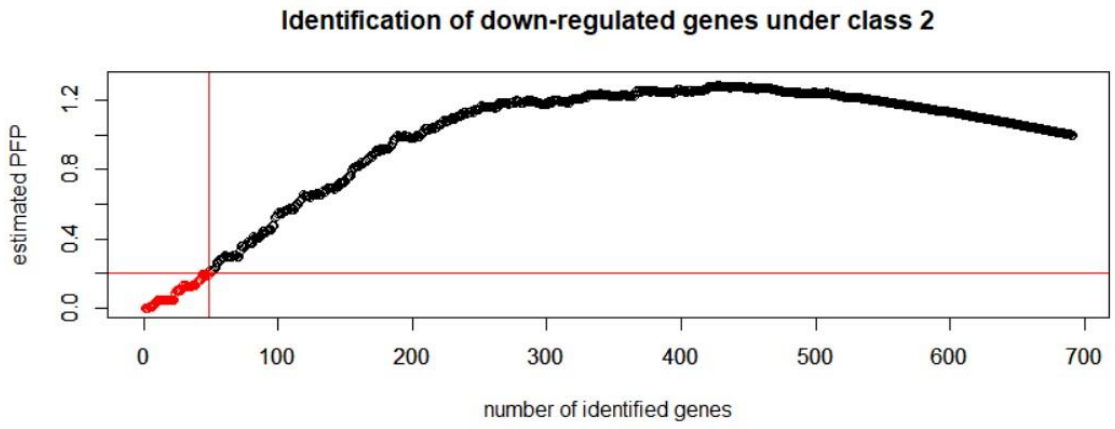

Figura A10 - Gráfico de identificação de massas reprimidas com FDR<0.2 para Fase Orgânica Positivo

Fase Orgânica Negativo com ESI(+):

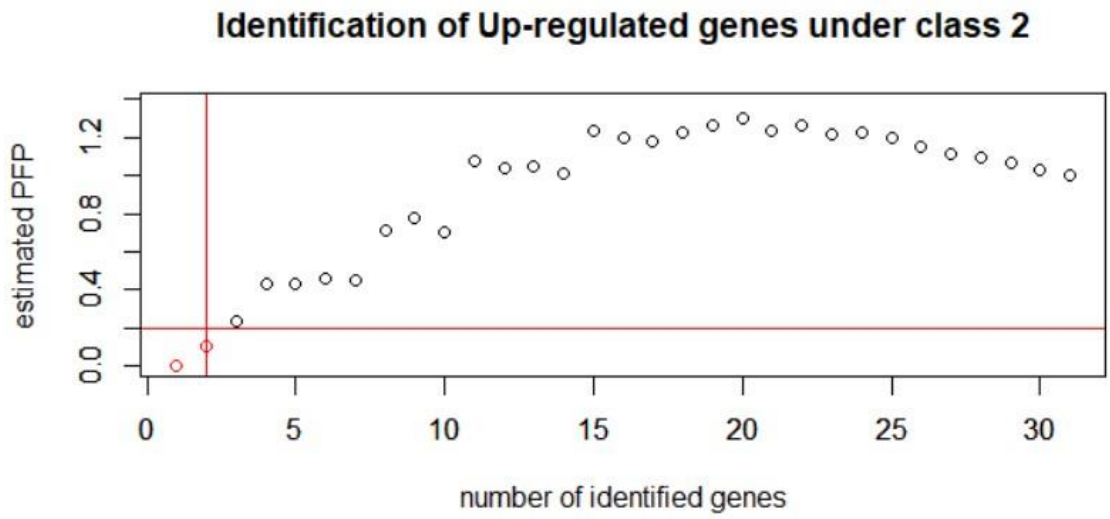

Figura A11 - Gráfico de identificação de massas induzidas com FDR<0.2 para Fase Orgânica Negativo

- Fase Orgânica Negativo com ESI(-):

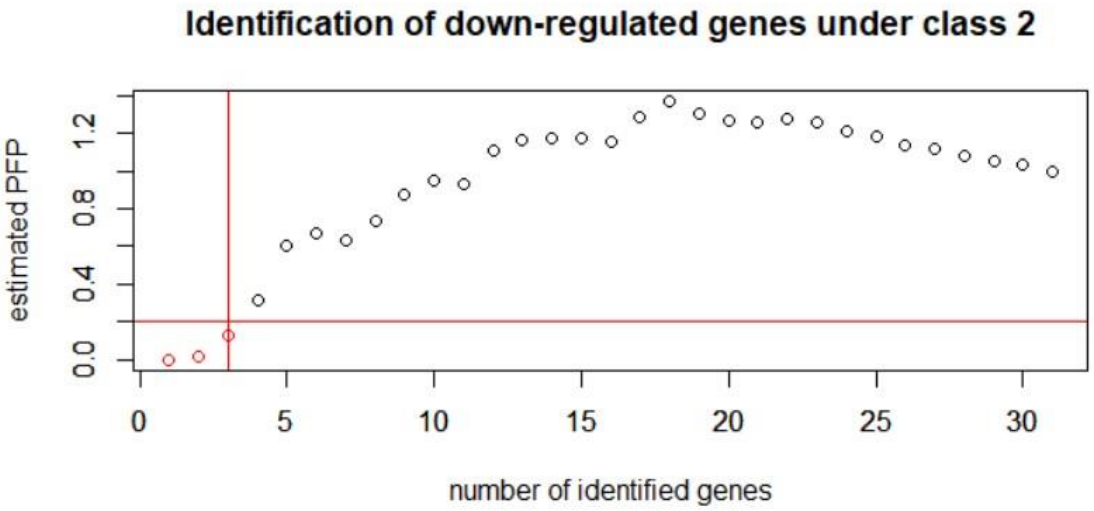

Figura A12 - Gráfico de identificação de massas reprimidas com FDR<0.2 para Fase Orgânica Negativo